UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

Experimental De

Samples and Sen ànica

Tesi Doctoral

UNIVERSITAT ROVIRA I VIRGILI

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

Experimental Design Applied to the Selection of Samples and Sensors in Multivariate Calibration

Tesi Doctoral

UNIVERSITAT ROVIRA I VIRGILI

UNIVERSITAT ROVIRA I VIRGILI

Departament de Química Analítica i Química Orgànica

Àrea de Química Analítica

EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION

Memòria presentada per

JOAN FERRÉ BALDRICH

per assolir el grau de

Doctor en Ciències Químiques

Tarragona, 1997

UNIVERSITAT ROVIRA I VIRGILI

DEPARTAMENT DE QUÍMICA ANALÍTICA
I QUÍMICA ORGÀNICA

Plaça Imperial Tàrraco, 1
43005 Tarragona
Tel. 34 (9) 77 55 81 37
Fax 34 (9) 77 55 95 63
e-mail: secqaqo@quimica.urv.es

Dr. FRANCESC XAVIER RIUS i FERRÚS, Catedràtic del Departament de Química Analítica i Química Orgànica de la Facultat de Química de la Universitat Rovira i Virgili,

CERTIFICA: Que la present memòria que duu per títol: "**EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION**", ha estat realitzada per en JOAN FERRÉ BALDRICH sota la meva direcció a l'Àrea de Química Analítica del Departament de Química Analítica i Química Orgànica d'aquesta Universitat i que tots els resultats presentats són fruit de les experiències realitzades per l'esmentat doctorant.

Tarragona, octubre de 1997

Prof. F. Xavier Rius i Ferrús

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

Als fulls que segueixen es resumeixen quasi quatre anys de feina davant d'un ordinador. Havia pensat paraules d'agraïment per a tots els que, d'una forma o altra, han estat per aquí tot aquest temps. Alguns han estat força transcendents i segurament han deixat la seva empremta en mi. Però he decidit resumir. Els que busqueu més avall el vostre nom, encara que no el trobeu escrit segur que hi sou. Així, doncs, voldria expressar la meva profunda gratitud ...

... en primer lloc al professor F. Xavier Rius, per haver acceptat dirigir aquest treball, per la seva dedicació i recolzament, per haver la seva confiança i ànims quan les coses no sortien i pel molt que m'ha ensenyat en tots els aspectes (no només científics) i que m'agradaria que continués ensenyant-me.

... al professor Roger Phan-Tan-Luu de la Université d'Aix-Marseille III, a qui considero un professor i una persona excepcional, per haver-me tractat sempre tant bé i haver acceptat estar al tribunal que jutja aquesta tesi. Ell em va transmetre la seva passió per la metodologia de la recerca experimental i a les seves ensenyances dec una part important d'aquest treball. Desitjo haver-les sabut reflectir prou bé.

... als doctors D.L. Massart, Romà Tauler, Jaume Puy i Itziar Ruisánchez, per haver-me fet l'honor d'acceptar jutjar aquest treball.

... als molts i bons companys i companyes, del grup de Quimiometria i de fora d'ell, pels molts moments que hem compartit aquests anys. I a algú molt especial per a mi... a tu, Montse (tu i jo ja sabem perquè).

... a tots aquells i aquelles de l'Àrea de Química Analítica que han tingut alguna cosa a veure, de prop o de lluny, amb la realització d'aquest treball.

... i sobretot moltes gràcies als de casa, pel seu constant suport i tots els i tots els sacrificis que desinteressadament han fet per mi.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

Als de casa...

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

CONTENTS

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

Chapter 1

Introduction and Objectives

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

1.1 Introduction

The aim of the chapter is to present this thesis: the objectives, the structure of the thesis and the work reported. A short bibliographic revision is used to justify the objectives.

The present introductory chapter has been structured in several sections. Section §1.1 introduces the importance of multivariate calibration in the present chemical analysis. This section serves, together with the bibliographic revision in §1.2, as a background to situate this thesis and its objectives in the context of the chemical analysis. The section §1.3 contains the objectives of this thesis and §1.4 presents the structure of this thesis. Finally, §1.5 contains the references cited in this chapter.

1.2 Multivariate calibration in quantitative analysis

Analytical chemistry plays an important role in our society. Chemical analyses can be performed for almost any substance of interest. Although many different problems of varied nature are presented to fulfill the needs of the present society, many situations in chemical analysis consist of identifying some constituents of a sample (qualitative analysis) or determining their concentration (quantitative analysis).

Quantitative analysis assumes that the measurands, usually concentrations of the constituents of interest in a sample, are related to the quantities measured from the technique used for analyzing the sample. These measured quantities can be a volume, a weight, or signals (e.g. spectra, chromatograms or voltages) from instruments such as spectrophotometers, chromatographs, potentiometers, etc.

In some procedures the analyst may only be interested in the raw numeric result of this measurement that is either compared to previous measured values

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

(e.g. in quality control) or used to detect relative changes in the data (e.g. inflexion points in a potentiometric curve). However, the usual situation when an instrumental technique is used to quantitatively analyze samples is to build a mathematical model relating the instrumental responses of a set of calibration samples to the quantities of chemical or physical variables such as analyte concentrations or indexes such as the octane number in fuels or the impact index in polymers. This relationship is used to predict these quantities from the instrumental response data of new *unknown* samples* measured in the same manner (e.g. spectra of test mixtures). Calibration is the process of building that mathematical model.

The motivation underlying multivariate calibration is to relate two types of measurements in a sample under study: one is easy to obtain and the other may either require expensive equipment, be time consuming, inaccurate or more difficult to obtain. The statistical model that establishes a relationship between the two series of measurements can be used for statistical inference of the unknown value of the difficult variable after observing the easy-obtainable variable. An example is the quantitative analysis using spectral data. Spectra are the *easy* measurements that can be related to the concentration of a determined analyte. Otherwise, the analyte should be obtained with the more costly reference or well-established method. If the instrumental method is faster and less costly than the reference method we could save time and money.

A variety of mathematical methods can be used to establish appropriate relationships between instrumental responses and chemical or physical measurands. Quantitative analysis has traditionally used univariate calibration based on a single measured signal and converting it to concentration via a calibration line. Two examples are the measurement of the pH and the single element atomic absorption. The major difficulty of this model is that the instrumental response must depend only on the concentration of the analyte of interest and all selectivity problems must be removed before the measurement. In these cases the analysis of mixture samples requires either a method to separate the analyte from the interferents or using a

* *Unknown sample* [1-5] will be used in this thesis to designate a sample to be analyzed (the problem sample). The usually objective is to predict the analyte concentration in this sample using a calibration model. The recommended name *test sample*[6] is not used to avoid confusion with a sample from the *test set* used in the modeling stage (see chapter 2).

highly selective instrument. The drawback is that the sample manipulation in the laboratory increases the cost of the analysis. On the other hand, the cost for making additional measurements on the unknown sample is decreasing due to the availability of analytical instruments that can perform many measurements per sample (multivariate signals) such as diode-array spectrophotometers and mass spectrometers. Hence, there is a tendency to make less sample manipulation and fewer experiments but to obtain more data in each of them. The simultaneous use of multiple responses and multivariate calibration can overcome the limitations of univariate calibration employed on methods which give *single point* measurements. Since many measurements are made in each sample, multivariate calibration can separate analytes from interferences without the need of highly selective measurements for the analyte and enables concentration determinations from non-selective measurements, thus improving the applicability of quantitative spectral analysis. The so called *inverse models* can calibrate for individual constituents in samples with very complex compositions, provided that the future unknown samples exhibit the same behavior as the calibration samples.

One advantage of multivariate signals combined with chemometric techniques is the considerable reduction in cost and time of the analysis. This is due to the reduction in the number of steps in the sample manipulation since it is not necessary to achieve complete separation or selectivity. Moreover, the concentration of more than one analyte can be determined at a time. However, more complex mathematical expressions than the simple univariate calibration are required. Sometimes, these models give a less precise or accurate result than the obtained with the traditional method of analysis but more rapid and cheaper. Examples of that are the analysis of protein in wheat or of water in meat. For these reasons, multivariate calibration methods are increasingly used in laboratories and spectrometers are becoming measurement devices of preferred choice in many current applications. Several reviews and textbooks are available on the subject[1,3,4,7-12].

The quality of the result of an analysis is influenced by all the steps involved in the analysis. Since prediction with multivariate calibration models is becoming a common step of the analytical procedure, the necessity of building these models that offer the guarantee of precise and unbiased predictions is clear. Actually, the development and improvement of these models is one of the focuses of chemometric research at the moment. Recent examples of this development are

new regression methods[13,14] and new equations that enable to understand the effect of the measurement errors and their propagation through the model on the error of the predicted concentrations[5]. The selection of the most adequate samples and sensors for calibration is also of primary concern for economical and statistical reasons. The economical reasons involve the cost of analyzing the calibration samples with a well-established or reference method and the cost of the instrumentation to measure the responses used in the analysis. The statistical reasons involve the ability of the quantitative calibration model to ensure good predictions for new samples. This is significantly influenced by the samples and sensors used for calibration.

Although the selection of sensors is regularly investigated, the proper selection of calibration samples seems more forgotten in the analytical literature and not many papers refer to this problem (see §3.1.3 for a bibliographic revision). In addition, the methods proposed for sample selection have some drawback about their real applicability such as not a clear mathematical criteria to decide the optimal number of samples to be used and to discard the samples that simply have redundant information. New criteria and methodologies must be developed to meet the quality criteria in multivariate calibration methods. The solution could be offered by the experimental design theory, until now used in multiple linear regression (MLR) but rarely employed in multivariate calibration models such as principal component regression (PCR) or partial least squares (PLS) regression.

Concerning the wavelength selection in multivariate calibration, the usual problem is how to identify the best range of wavelengths or individual sensors for prediction. Different methods and criteria for optimal wavelength selection have been proposed, specially in classical least squares regression (CLS) and the experimenter interested in using a selection criterion must review (and understand!) a large volume of literature. In addition, there is some discrepancies about the performance of these criteria and whether if they improve the precision, the trueness or the accuracy of the result. Many times the authors do no define what they consider to be accuracy or precision so that their results are difficult to compare. The usefulness of these criteria and their effect on the modern concepts of precision and trueness of the results should be clarified.

Moreover, a software to calculate the criteria and methodologies should be elaborated to facilitate its application in the laboratory.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

1.3 Objectives of the thesis

The objective of this doctoral thesis is to study the sample and sensor selection criteria for the optimization of the multivariate calibration models used in analytical methods using the concepts from the experimental design. The criteria found in the scientific literature are considered and, when possible, improved procedures for sensor and sample selection are proposed.

More specifically, we contribute to the study of:

1. A new sample selection method in principal components regression (PCR) based on the application of the D-optimality criterion and the Fedorov's exchange algorithm. The method only uses the instrumental responses of the candidate samples to select the minimum number of calibration samples to build the model. The analyte concentration is only determined for the selected samples. Complementing this method, a new procedure for the fast selection of the relevant factors in PCR is devised. The Fedorov's algorithm was also applied to select samples to check if model standardization is necessary in PCR models. The performance of the sets selected using the D-criterion or the Kennard-Stone algorithm in MLR was compared.

2. New guidelines for sensor selection in CLS based on the experimental design theory. The different criteria for wavelength selection are critically reviewed in the modern terms of the precision , accuracy and trueness. They are interpreted from the point of view of the experimental design theory using the confidence hyperellipsoid of the predicted concentrations. The effects of collinearity in CLS and its effects in the quality of the predicted concentrations have been studied.

3. A new method to detect and reduce bias in future test samples in multicomponent analysis (CLS) .

4. The methodologies and algorithms have been written in m-files Matlab. In the future they will be implemented in Toolbox for their application in the laboratory.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

1.4 Structure of the work presented

The thesis has been structured in five chapters. Each chapter is divided in sections. The first section of each chapter is the introduction and contains the aim, the summary of the contents of the chapter and the justification of the objectives with a bibliographic revision. This section ends with the references used in the bibliographic revision. The next sections in the chapter contain the experimental work written as papers (either published, submitted or in preparation). Each paper contains the following parts: introduction, theoretical background, experimental part, discussion and results, conclusions and references. The conclusions of each chapter are, together with the general conclusions of the thesis, in the chapter 5. The contents of each chapter are the following:

- Chapter 1 contains the introduction to the work reported in this thesis, the structure of the thesis and the major aims of the work that is developed in the following chapters.

- Chapter 2 contains six sections of theoretical (not experimental) nature. Section §2.1 is the introduction to the chapter. Section §2.2 has the notation and list of symbols used in this thesis. The notation used in the published papers is adequately indicated in the paper and is very close to the indicated in §2.2. Section §2.3 introduces the multiple linear regression (MLR) and the least squares solution, quite used in this thesis. Some concepts of the experimental design theory, focused on the optimality criteria to select samples in MLR are reviewed. These ideas are then used in the chapters §3 and §4. Section §2.4 deals with the basic concepts of the multivariate calibration methods used in this thesis. A review of their theoretical basis, that is disperse in many papers and monographs, is considered important to understand the need of sample and wavelength selection. This part also presents some new concepts related to the net analyte signal in classical least-squares (CLS) calibration that are later used in the section §4.3 and §4.7. The section §2.5 contains guidelines for applying experimental design to the calibration models. These ideas are used for sample and wavelength selection in chapters 3 and 4. Chapter 2 ends with the references (§2.6) of this chapter.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

• Chapter 3 deals mainly with the selection of the best calibration sample subset for PCR. The procedures proposed in the literature for sample selection are critically reviewed (§3.1.3) A new methodology to select samples from the instrumental responses of a large subset when the samples cannot be synthesized is presented in *Selection of best calibration sample subset for multivariate regression.* Joan Ferré, F. Xavier Rius *Anal. Chem.* 68, (1996) 1565-1571 (section §3.2). The next paper *Determination of ethylene content in poly(propylene-ethylene) copolymers using near-infrared spectra (NIR) and multivariate calibration* Villagrasa C., Ferré J., Larrechi M.S., Rius F.X., García C. (*in preparation*) (section §3.3) compares the predictive ability of PLS and of PCR with the factors selected according to the methodology described in §3.2 using near-infrared (NIR) data of industrial copolymers. Section §3.4 is the paper *Constructing D-optimal designs from a list of candidate samples.* Joan Ferré, F. Xavier Rius *Trends Anal. Chem.* 16 (1997) 70-73 and is a comparative study of the Fedorov's algorithm, the popular Kennard-Stone algorithm and the random division of samples into calibration and validation sets for the selection of samples for a MLR model. §3.5 is the paper *Selection of calibration points for PCR in QSAR studies.* Joan Ferré, F. X. Rius (*in preparation*). It deals with the selection of calibration samples in quantitative structure-activity relationship (QSAR) studies based on the principal components using the Fedorov's algorithm. The last paper, *Assessing the validity of principal component regression models in different analytical conditions.* Rius A.; Callao M.P., Ferré J.; Rius F.X., *Anal. Chim. Acta* 337 (1997) 287-296 presents a procedure for selecting samples for assessing if a PCR model is still valid before using the piecewise direct standardization (PDS) technique when the working conditions are different from those used for modeling. The contribution to this work consists on applying the D-optimality criterion for selecting, from a large set, the minimum number of samples that must be analyzed in the new conditions.

• Chapter 4 deals with the selection of the best sensor subset for multicomponent analysis. In the paper *A graphical criterion to examine the quality of multicomponent analysis. Implications for wavelength selection.* J. Ferré and F.X. Rius *Trends Anal. Chem.* 16 (1997) 155-162 (section §4.2) the basic concepts used in sensor selection in CLS are explained on the basis of their effect on the volume, shape and orientation of the confidence region (an ellipsoid) of the predicted concentrations. This paper is complemented with *Further considerations on the sensitivity and selectivity of multicomponent systems.* J. Ferré and F.X. Rius. *In preparation* (section §4.3). Here, the

mathematical expressions of sensitivity, selectivity, variance-proportion decompositions and condition number are discussed and interpreted using the confidence ellipsoid. The effect of adding a new sensor to the calibration matrix on the confidence ellipsoid is shown. The section §4.4 is the paper *Equivalence between Selectivity and Variance Inflation factors in multicomponent analysis.* J.Ferré, F.X. Rius *Química Analítica* 15 (1996) 259-262, where the mathematical equivalence between selectivity and variance inflation factors is shown. Both can be used as measures of collinearity in CLS. The section §4.5 is a tutorial where the ISO definitions of accuracy, trueness and precision are reviewed and their relationships with the wavelength selection criteria in CLS are revisited and clarified. This is motivated by the confusion in the literature about the effect of these criteria on the accuracy, trueness and precision of the results. The section §4.6 is the paper *Figures of Merit in Multivariate Calibration. Determination of Four Pesticides in Water by FIA and Spectrophotometric Detection.* J. Ferré, R. Boqué, B. Fernández-Band, M.S. Larrechi and F.X. Rius. *Anal. Chim. Acta* 348 (1997) 167-175. This paper studies the variance proportion decompositions and the effects of the selectivity and sensitivity in the prediction error of four pesticides analyzed with a FIA system and CLS. The whole published paper has been included although the part concerning the detection limits is not a subject of this thesis. The last part of chapter 4, §4.7, is the paper *Detection and correction of biased results of individual analytes in multicomponent spectroscopic analysis* J.Ferré, F.X. Rius, *Submitted for publication.* This work was motivated by the fact that, in the paper in §4.6, the large selectivity and sensitivity values of the analytes did not agree with the calculated prediction errors. It was supposed that the large inexplicable errors were not due to the instability of the system of equations but to an erroneous preparation of the validation samples with a deficient assigned value of the concentration. In the present paper a tool for internal validation of the standards and the validation samples based on the net analyte signal is developed. It enables the bias in CLS models to be detected taking advantage of the multivariate signal. A wavelength selection procedure is used to select the wavelength with less prediction error.

• Chapter 5 contains the conclusions of the chapters 3 and 4 and the general conclusions of the thesis. The advantages and drawbacks of the proposed methodologies are commented and the trends for future work are devised. It must be reminded that each paper already has its particular conclusions.

The format of the published papers in this thesis. The published papers have been edited to give a uniform format to the thesis but the contents have not been changed. The only change in nomenclature with respect to what has been published can be found in paper §4.6 *Equivalence between Selectivity and Variance Inflation factors in multicomponent analysis.* J.Ferré, F.X. Rius *Química Analítica* 15 (1996) 259-262 where the sign ' , used in the paper to indicate transposition, has been changed in the thesis for a ^{T}.

1.5 References

1. Haaland D.M., Thomas E.V. *Anal. Chem.* 60 (1988) 1193-1202.
2. Lorber A., Kowalski B.R. *J. Chemom.* 2 (1988) 93-109.
3. Booksh K.S., Kowalski B.R *Anal. Chem.* 66 (1994) 782A-804A.
4. Beebe K.R., Kowalski B.R. *Anal. Chem.* 59 (1987) 1007A-1017A.
5. Faber K., Kowalski B.R. *J. Chemom.* 11 (1997) 181-238.
6. ISO 3534-1 (1993) *Statistics-Vocabulary and symbols.* International Organization for Standardization, Geneva, Switzerland.
7. Martens H., Naes T. *Multivariate Calibration,* Wiley: New York, 1989.
8. Sanchez E., Kowalski B.R. *J. Chemom.* 2 (1988) 247-263.
9. Kowalski B.R., Seasholtz M.B. *J. Chemom.* 5 (1991)129-145.
10. Thomas E.V. *Anal. Chem.* 66 (1994) 795A-804A.
11. Geladi P., Kowalski B.R. *Anal. Chim. Acta.* 185 (1986) 1-17.
12. Martens H., Karstang T., Naes T. *J. Chemom.* 1 (1987) 201-219.
13. Vigneau E., Devaux M.F., Qannari E.M., Robert P. *J. Chemom.* 11 (1997) 239-249.
14. Wentzell P.D., Andrews D.T., Kowalski B.R. *Anal. Chem.* 69 (1997) 2299-2311.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

2.1. Introduction

2.1.1 Aim of the chapter

The aim of this chapter is to introduce the concepts about experimental design and multivariate calibration used in this thesis. It is also a compendium of the large amount of information disperse in many papers and books.

2.1.2 Structure of the chapter

This chapter has six parts: introduction (§2.1), notation and definitions (§2.2), concepts of experimental design theory in general statistical terms (§2.3), the theoretical background of the multivariate calibration models used in this thesis and their advantages and limitations (§2.4), the connection points between the experimental design theory and the multivariate calibration models (§2.5) and the collinearity problem in the calibration models (§2.6).

2.2 Notation and definitions

Multivariate calibration can be used with any type of suitable multivariate data to model any property. However, the terminology used here associates instrumental responses with sample spectra and the property of interest with the analyte concentration since the major part of this thesis deals with this type of data.

Matrices are represented by bold capital letters, e.g. **R**, bold lowercase letters denote column vectors, e.g. **c** (row vectors are transposed column vectors) and italic characters represent scalars, e.g. c_k. True values are indicated by Greek characters or

the subscript $_{true}$. Calculated or measured values are indicated by Roman characters. The *hat* (^), used in the literature to indicate *calculated,* has been dropped from the symbols to simplify the notation; if the magnitude is measured or calculated can be deduced from the context. The running indexes in multivariate calibration are: $k = 1$ to K analytes are present in $i = 1$ to I calibration samples whose instrumental responses are measured using j=1 to J sensors. The following is the list of symbols used in this thesis:

Symbols beginning with a Roman letter

A-	criterion of the trace of the dispersion matrix
D-	criterion of the determinant of the dispersion matrix
G-	criterion of the maximal variance function
a	index for factors
A	dimensionality (number of factors)of PCR or PLS models
$\mathbf{a}_k=[a_{1,k},\ldots\ a_{j,k}\ \ldots\ a_{J,K}]^T$	(J×1) spectrum of the analyte k at J wavelengths at concentration c_k^0
$\mathbf{a}_k^*$	(J×1) net analyte signal for the analyte k in a pure analyte spectrum at concentration c_k^0
$\mathbf{A}=[\mathbf{a}_1,\ldots\ \mathbf{a}_k\ \ldots\ \mathbf{a}_K]$	(J×K) spectra of the K components at J wavelengths at concentration c_k^0
$\mathbf{A}_k=[\mathbf{a}_1,\ldots,\ \mathbf{a}_{k-1},\ \mathbf{a}_{k+1},\ldots\mathbf{a}_K]$	(J×(K-1)) matrix **A** without the *kth* column $\mathbf{a}_k$
b_j	estimated coefficient of the variable j
$b_{j,k}$	estimated coefficient of the variable j for the analyte k
$\mathbf{b}=[(b_0),\ b_1,\ \ldots,\ b_j,\ \ldots,\ b_J]^T$	((J+1)×1 or J×1) estimated coefficients
$\mathbf{b}_k=[b_{1,k},\ \ldots,\ b_{j,k},\ \ldots,\ b_{J,k}]^T$	(J×1) estimated coefficients for the analyte k
$\mathbf{b}_{k,CLS}\ \mathbf{b}_{k,ILS}\ \mathbf{b}_{k,PCR}\ \mathbf{b}_{k,PLS}$	vectors of coefficients for the analyte k estimated in the CLS, ILS, PCR and PLS models respectively
$\mathbf{B}=\{b_{j,k}\}=[\mathbf{b}_1,\ \ldots,\ \mathbf{b}_k,\ \ldots,\ \mathbf{b}_K]$	(J×K) estimated coefficients for K analytes and J variables
$c_{i,k}$	concentration of the analyte k in the sample i
$c_{un,k}$	concentration of the analyte k in the unknown sample
$\overline{c}_k$	mean value of $\mathbf{c}_k$
$\mathbf{c}_k=[c_{1,k},\ \ldots,\ c_{i,k},\ \ldots,\ c_{I,k}]^T$	(I×1) concentration of the analyte k in I calibration samples
$\mathbf{c}$	(K×1) concentration of K analytes in a sample
$\mathbf{c}_{un}$	(K×1) concentration of K analytes in an unknown sample
$\mathbf{C}=\{c_{i,k}\}=[\mathbf{c}_1,\ \ldots,\ \mathbf{c}_k,\ \ldots,\ \mathbf{c}_K]$	(I×K) concentrations of K analytes in I calibration samples. A column is the concentration of one analyte in the I calibration samples. A row is the concentration of the K analytes in one

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

	calibration sample
$\mathbf{C}_0$	(K×K) diagonal matrix whose diagonal elements are c_k^0
$cov(x_1,x_2)$	covariance of the random variables x_1 and x_2
$d(\mathbf{x}_{un})$	variance function at the point $\mathbf{x}_{un}$
$\mathbf{D}$	(dimensions depend on the equation) diagonal matrix of singular values of **R** or **S**
$\mathbf{D}_A$	(A×A) diagonal matrix of A singular values of **R** or **S**
$\mathbf{E}$	(dimensions depend on the equation) residual matrix
$\mathbf{f}(\mathbf{x}^i)$	(Q×1) regression functions evaluated at the point $\mathbf{x}^i$. Transposed row of the **X** matrix
$F_{\nu 1,\ \nu 2,\alpha}$	upper α-percentage point of F-distribution with ν_1 and ν_2 degrees of freedom
$\mathbf{H}$	(I×I) orthogonal projection matrix. Hat matrix.
i	index for experiments, calibration samples, objects or points
I	number of experiments or samples in the calibration matrix
I_P	number of samples in the validation set
$\mathbf{I}$	appropriately dimensioned identity matrix
j	index for independent variables or wavelengths or sensors
J	number of independent variables or wavelengths in a spectrum
k	index for analytes
K	number of analytes or components or constituents in a sample
$\mathbf{M}$	(Q×Q) normalized information matrix or matrix of moments
$\max(\lambda_j)$	the largest eigenvalue of $(\mathbf{X}^T\mathbf{X})^{-1}$
$\min(\lambda_j)$	the lowest eigenvalues of $(\mathbf{X}^T\mathbf{X})^{-1}$
N	number of candidate points
$N(\mu,\sigma^2)$	normal distribution with mean μ and variance σ^2
$\mathbf{P}$	(J×J) eigenvectors of $\mathbf{R}^T\mathbf{R}$
$\mathbf{p}_a$	(J×1) ath column of **P**, eigenvector of the ath factor
$\mathbf{P}_A$	(J×A) selected factors
Q	number of coefficients in a MLR model
r_j	instrumental response measured at the sensor j
$r_{i,j}$	instrumental response of the sample i measured at the sensor j
$r_{un,j}$	instrumental response of the unknown sample measured at the sensor j
$r_{un,j,k}^*$	net analyte signal of the analyte k at the sensor j in the unknown sample
$\mathbf{r}=[r_1, ..., r_j, ..., r_J]^T$	(J×1) instrumental responses measured at J sensors (e.g. absorbances at J wavelengths)
$\mathbf{r}_i=[r_{i,1}, ..., r_{i,j}, ..., r_{i,J}]^T$	(J×1) instrumental responses of the calibration sample i measured at J sensors
$\mathbf{r}_{un}=[r_{un,1}, ..., r_{un,j}, ..., r_{un,J}]^T$	(J×1) instrumental responses of the unknown sample measured at J sensors

$\mathbf{r}_{un,k}^*$	(J×1) net analyte signal for the analyte k in the unknown sample
$\bar{\mathbf{r}}$	(J×1) column means of $\mathbf{R}$
$\mathbf{R}=\{r_{i,j}\}$	(I×J) instrumental responses of I calibration samples measured at J sensors
s^2	estimate of σ^2
$s_{j,k}$	partial sensitivity of the analyte k in the sensor j, defined as the slope of the analytical calibration plot of the response of the sensor j to the concentration of the analyte k (responses divided by the concentration of the analyte in a pure sample). It is usually the molar absorptivity coefficient of the pure component at unit pathlenght
$s_{j,k}^*$	net analyte signal of the analyte k in the sensor j of the spectrum of the analyte k pure
$\mathbf{s}_k=[s_{1,k},\ldots s_{j,k} \ldots s_{J,K}]^T$	(J×1) partial sensitivities of the analyte k at J wavelengths
$\mathbf{s}_k^*$	(J×1) net analyte signal for the analyte k in a pure analyte spectrum
$\mathbf{S}=[\mathbf{s}_1,\ldots \mathbf{s}_k \ldots \mathbf{s}_K]$	(J×K) partial sensitivities of the K components at J wavelengths
$\mathbf{S}_k=[\mathbf{s}_1,\ldots, \mathbf{s}_{k-1}, \mathbf{s}_{k+1},\ldots \mathbf{s}_K]$	(J×(K-1)) matrix $\mathbf{S}$ *without the kth column* $\mathbf{s}_k$
$t_{i,a}$	ath score associated with the ith sample
$\mathbf{t}_a$	(I×1) sample scores for the ath factor
$\mathbf{t}_A^i$	(A×1) scores of the sample i for A factors
$\mathbf{t}_{un,A}$	(A×1) scores of the unknown sample for A factors
$\mathbf{T}_A$	(I×A) scores of $\mathbf{R}$ for A factors
$\mathbf{U}_A$	(I×A) normalized PCA scores for A factors
$UVIF_j$	jth diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$ called the unscaled variance inflation factor (or just variance coefficient) of the coefficient j
$var()$	variance of a scalar quantity
$\mathrm{var}(\mathbf{v})$	symmetric variance-covariance matrix of the vector $\mathbf{v}$
$\mathbf{v}_a$	ath eigenvector of $(\mathbf{X}^T\mathbf{X})^{-1}$
VIF_j	variance inflation factor of the coefficient j
x	independent variable or input variable
$x_{i,j}$	value of the variable j in the experiment (point) i
$\mathbf{x}_j$	column j of $\mathbf{X}$
$\mathbf{x}^i$	row i of $\mathbf{X}$, a point of the experimental domain
$\bar{\mathbf{x}}$	column means of $\mathbf{X}_{-1}$
$\mathbf{X}=(\mathbf{x}_j)=(\mathbf{x}^i)$	(I×Q) (usually I×J or I×J+1) model matrix
$\mathbf{X}_{-1}$	(I×Q-1) model matrix with the column of ones deleted
$\mathbf{X}^T\mathbf{X}$	(Q×Q) information matrix
$(\mathbf{X}^T\mathbf{X})^{-1}$	(Q×Q) dispersion matrix
$(\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$	(Q×Q) variance-covariance matrix
y	dependent variable, output variable, response
y_i	measured output variable in the experiment i

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

y_{un}	predicted output variable in the experiment the point x_{un}
$\bar{y}$	mean value of the elements of **y**
y	(I×1) measured output variable in I experiments
$\mathbf{y}_{pred}$	(I×1) predicted values of the output variable in I experiments

Symbols beginning with a Greek letter

α	significance level of a statistical test
β_j	true regression coefficient for the variable j
$\beta=[(\beta_0), \beta_1,...,\beta_j,...,\beta_J]^T$	(J×1 or (J+1)×1) true regression coefficients of the model
$\beta_k=[\beta_{1,k},...,\beta_{j,k},...,\beta_{J,k}]^T$	(J×1) true regression coefficients for the analyte k
ε_i	unknown error in the experiment i
ε	(I×1) unknown errors in I experiments
η_i	true value of the output variable in the experiment i
η	(I×1) true values of the output variable in I experiments
λ_a	ath eigenvalue of $(\mathbf{X}^T\mathbf{X})^{-1}$
λ_j	jth eigenvalue of $(\mathbf{S}^T\mathbf{S})^{-1}$
σ_a	ath singular value of **R**
σ_i	ith singular value of **S**
σ	standard deviation
σ^2	variance of a scalar quantity
ξ_N	(N×J) matrix of candidate points
ξ_I	(I×J) matrix of experiments (points)
ξ^{*C}	matrix of experiments optimal for the C-criterion
χ	experimental domain of interest
Ξ_I	set of matrices of I experiments (points)
$\theta_{k,a}$	true regression coefficient for scores of factor a
θ_k	true regression vector with respect to scores

Other symbols

, (coma)	adds columns in matrices or vectors: $\mathbf{P}=[\mathbf{p}_1, \mathbf{p}_2]=[\mathbf{p}_1\ \mathbf{p}_2]$
; (semicolon)	adds rows in matrices or vectors: $\mathbf{X}=[\mathbf{x}_1^T; \mathbf{x}_2^T] = \begin{bmatrix}\mathbf{x}_1^T \\ \mathbf{x}_2^T\end{bmatrix}$
*	(superscript) net analyte signal
$\|\cdot\|$	2-norm: Euclidean norm
1	appropriately dimensioned vector of ones
Cond(**X**)	condition number of the matrix **X**
Det(**X**)	determinant of the matrix **X**

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

$E[\cdot]$	expected value
eq eqs	equation equations
$i \times j$	(subscript) dimensions of a vector or matrix
k-row k-col	(subscript) kth row or column of the matrix
max	maximum value in a list
min	minimum value in a list
Tr(X)	trace of the matrix **X**
$\mathbf{X}^T$ $\mathbf{x}^T$	(superscript) transposition of a matrix **X** or a vector **x**
$\mathbf{X}^+$	(superscript) Moore-Penrose pseudo-inverse of the matrix **X**
$\mathbf{X}^{-1}$	(superscript) the inverse of a matrix of the matrix **X**
un	(subscript) magnitude related to the unknown sample

Definitions

Orthonormal. If **X** is orthonormal then $\mathbf{X}^T\mathbf{X}=\mathbf{X}\mathbf{X}^T=\mathbf{I}$

Singular-value decomposition (SVD). Decomposes a matrix **X** ($I \times Q$ of rank k) into three matrices $\mathbf{X}=\mathbf{U}\mathbf{D}\mathbf{P}^T$ with $\mathbf{U}^T\mathbf{U}=\mathbf{P}^T\mathbf{P}=\mathbf{I}$ ($k \times k$) and **D** diagonal with the k positive singular values on the diagonal.

Moore-Penrose pseudo-inverse [1,2]: used for non-square or overdetermined matrices, where the inverse cannot be calculated. The pseudo-inverse of a matrix **X** ($I \times Q$ of rank k) is calculated as $\mathbf{X}^+=\mathbf{P}\mathbf{D}^{-1}\mathbf{U}^T$ where **P**, **U** and **D** result from the SVD of **X**. The pseudo-inverse is also $\mathbf{X}^+=(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ if **X** has linearly independent columns. If **y** is an I-vector of responses, the ordinary least squares (OLS) estimator of the regression coefficients is $\mathbf{b}_{OLS}=(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}=\mathbf{X}^+\mathbf{y}$. When some rows or columns of **X** are linearly dependent or **X** has more columns than rows (i.e. **X** is not of full rank, $k<Q$), the determinant of $\mathbf{X}^T\mathbf{X}$ is zero, $(\mathbf{X}^T\mathbf{X})^{-1}$ does not exist and there is no unique least squares estimator. However, the pseudo-inverse can still be calculated and $\mathbf{b}_{MLLS}=\mathbf{X}^+\mathbf{y}$ is the unique minimum length least squares (MLLS) solution, which means that the solution **b** minimizes $\mathbf{b}^T\mathbf{b}$ in the case that $\mathbf{X}^T\mathbf{X}$ is singular[3]. The OLS and MLLS estimator coincide in the full rank case[4].

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

2.3 Experimental design in multiple linear regression

2.3.1 The study of chemical systems

Chemical systems can be represented as in Figure 2.1 where one or more outputs y, also called *dependent variables* or *responses* can be measured. Their result depends of one or more input variables x *, also called *independent variables* [5]:

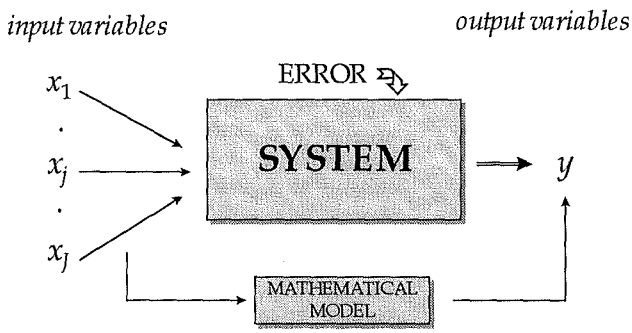


Figure 2.1. Representation of a system under study.

The aim of research in many chemistry areas is to understand and improve the system under study by finding the dependence between the input and the output variables. This requires conducting *experiments* by varying the values (*levels*) of one or more input variables and studying the changes in the response. The comparison of the obtained data puts into evidence that dependence and enables to identify the optimal conditions, the variables that most influence the results and those that do not, the presence of interactions, etc. The range of values that can take each variable is the *domain* and the combination of the domain of all the variables is the *domain of the variables.* This contains all the feasible experiments (the *possible experimental domain*) from where the *experimental domain of interest* (χ) may be a part

* The word *factor* used in experimental design as the cause of the studied phenomenon is not used here to avoid confusion with the *factors* from PCA and PLS.

of it. The variables are usually expressed in coded units. The set of I experiments to be performed, expressed in coded units, constitutes the *matrix of experiments* (ξ_I), where each row represents one experiment and each column one of the J variables. The *experimental plan* is the matrix of experiments expressed in raw variables.

To study such dependences, the design of the experimental runs is many times based on the experimenter's intuition and it is not rare to use the sequential method of changing one-variable-at-a-time experiments. However, this has been shown to be inefficient[6-8]: it may require too many experiments if the number of variables and levels to study is large and it provides no information on interactions among the variables. A more efficient way of conducting the study consists of changing simultaneously several variables in the same experiment. The important decisions derived from the experimental results and the non-negligible cost of the experimentation may advise using a methodological approach to establish the optimal organization of the experiments. Statistical experimental designs provide the mathematical framework for studying several variables simultaneously in a small number of experiments and obtaining information on interaction behavior.

2.3.2 Aims of experimental design

Experimental design, also called *design of experiments* (DOE), stands for a systematic way of planning the experiments aimed at efficiently extracting information employing statistical tools. The DOE helps the experimenter to select an optimal experimental strategy to achieve the proposed objectives. This includes:

1. To devise a reduced-cost set of experiments necessary to obtain the answer to a problem (e.g. the influence of the variables on the system or optimization of a response through a mathematical model among others[9]) by varying systematically multiple variables in a single experiment. This ensures the maximum efficiency of the experimental work and avoids random assays, redundant information and the pitfalls of the one-component-at-a-time method. This affects the cost of the experimentation.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

2. To assure that the selected experiments will yield a correct and reliable information about the influence of each variable on the system by minimizing the variance of estimated coefficients obtained through regression[10]. In addition the experiments must enable to interpret the obtained information and the generalization of the conclusions in the domain of interest. Statistical analysis methods are employed to determine whether a treatment is statistically significant in influencing the system response. This affects the quality of the result.

2.3.3 The linear model of multiple independent variables

2.3.3.1 The multiple linear regression (MLR) model

Many times the dependece between the input and output variables is expressed as a *mathematical model* * :

$$\eta = F(x_1, \ldots x_j, \ldots x_J, \beta_1, \ldots \beta_q, \ldots \beta_Q) \qquad (2.1)$$

where η is the true value of the response and β_q (q=1, ..., Q) are the true coefficients of the model, supposed constant in the domain under study. The model can be used to describe experimental results, to interpret the influence of the independent variables in the response or for prediction purposes in the experimental domain. The mathematical expression depends on the phenomenon being described and on the domain (usually, the smaller the domain, the simpler the model can be). The models considered here are linear with respect to coefficients since they usually represent a sufficient approximation to the reality of the domain. They can be written as:

$$\eta = \mathbf{f}^T(x_1, \ldots x_j, \ldots x_J)\beta \qquad (2.2)$$

where β is the vector of coefficients and $\mathbf{f}(x_1, \ldots x_j, \ldots x_J)$ is a vector of x variables that represents the equation of the model (it may include, for example, the

* Since regression actually is a statistical problem, along §2.3 the common symbols x and y are used for the independent and dependent variables respectively instead of a specialized chemical notation.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

measured variables, cross-products and powers e.g. in $\eta=\beta_0 + \beta_1 x_1 + \beta_{11} x_1^2$, the vectors are $\mathbf{f}^T=[1\ x_1\ x_1^2]$ and $\beta=[\beta_0\ \beta_1\ \beta_{11}]^T$). The models considered in this thesis are first degree polynomia for J independent variables:

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_J x_J \qquad (2.3)$$

so that $\mathbf{f}^T=[1, x_1, x_2, ... x_J]$ and $\beta=[\beta_0, \beta_1, \beta_2, ..., \beta_J]^T$. Since some unknown error ε_i (random and/or systematic) is always present in any experimentally obtained quantity, the measured response value in the experiment i is not the true value η_i but

$$y_i = \eta_i + \varepsilon_i \qquad (2.4)$$

For this reason, the same experiment performed several times in conditions as similar as possible will never give the same response. Here ε is assumed to be independent (for two experiments i and i' $cov(\varepsilon_i, \varepsilon_{i'})=0$) and normally distributed with mean 0 and variance σ^2. Eq 2.4 is then written as:

$$y_i = \mathbf{f}^T(\mathbf{x}^i)\,\beta + \varepsilon_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + ... + \beta_J x_{i,J} + \varepsilon_i \qquad (2.5)$$

where $\mathbf{f}^T(\mathbf{x}^i)$ is the vector of the variable settings of the ith experiment $\mathbf{x}^i=[x_{i,1}, x_{i,2}, ... x_{i,J}]$. The equations necessary to estimate the coefficients are obtained by conducting the experiments described in the *matrix of experiments* ξ_I (where I must not be inferior to the number of coefficients in the model).The matrix notation for the I experiments is (Figure 2.2)

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon \qquad (2.6)$$

$$\underset{I}{\begin{bmatrix} y_1 \\ \cdot \\ y_i \\ \cdot \\ y_I \end{bmatrix}} = \underset{I \times Q}{\begin{bmatrix} 1 & x_{11} & \cdots & x_{1j} & \cdots & x_{1J} \\ \cdot & \cdot & \cdots & \cdot & \cdots & \cdot \\ 1 & x_{i1} & \cdots & x_{ij} & \cdots & x_{iJ} \\ \cdot & \cdot & \cdots & \cdot & \cdots & \cdot \\ 1 & x_{I1} & \cdots & x_{Ij} & \cdots & x_{IJ} \end{bmatrix}} \underset{Q}{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ \beta_j \\ \cdot \\ \beta_J \end{bmatrix}} + \underset{I}{\begin{bmatrix} \varepsilon_1 \\ \cdot \\ \varepsilon_i \\ \cdot \\ \varepsilon_I \end{bmatrix}}$$

Figure 2.2 Matrix representation on eq 2.6.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

where $\mathbf{y}_{I\times1}$ is the vector of observed responses, $\mathbf{X}_{I\times Q} = [\mathbf{f}^T(\mathbf{x}^1); ...; \mathbf{f}^T(\mathbf{x}^i); ... ; \mathbf{f}^T(\mathbf{x}^I)]$ is called the *model matrix* and $\varepsilon_{I\times1}$ is the vector of non-observable errors supposed $E(\varepsilon)=0$ and variance-covariance matrix $var(\varepsilon)=\sigma^2\mathbf{I}$. If the model is given by eq 2.5, the ith row of $\mathbf{X}_{I\times Q}$ is $\mathbf{f}^T(\mathbf{x}^i) =[1\ x_{i,1}, x_{i,2}, ... x_{i,J}]$.

2.3.3.2 Estimation of the model. The least-squares solution

Since ε in eq 2.6 is unknown, only an estimation **b** of β can be found. Different approaches[2] exit for estimatingβ. The least-squares solution is given by :

$$\mathbf{b}= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \qquad (2.7)$$

where $\mathbf{b}=[b_0, ... b_j, ... , b_J]^T$ is the vector of estimated regression coefficients. The properties of this solution depend on the validity of some requirements (see references 11,12). Notice that to evaluate $(\mathbf{X}^T\mathbf{X})^{-1}$ (the *dispersion matrix*) the experiments must be chosen in such a manner that $\mathbf{X}^T\mathbf{X}$ (the *information matrix*) is non-singular. If **X** is a full rank matrix, **b** in eq 2.7 can also be calculated as (see §2.2):

$$\mathbf{b}= \mathbf{X}^+\mathbf{y} \qquad (2.8)$$

By defining $\mathbf{X}_{-1}$ as the matrix **X** with the column of ones deleted ($\mathbf{X}=[\mathbf{1},\mathbf{X}_{-1}]$) and $\mathbf{X}_c$ as $\mathbf{X}_{-1}$ after column-centering , the same solution can be found using column-centered data [2 page 369, 12 page 192, 13 page 43]:

$$\mathbf{b}_1= (\mathbf{X}_c^T\mathbf{X}_c)^{-1}\mathbf{X}_c^T\mathbf{y}_c = (\mathbf{X}_c^T\mathbf{X}_c)^{-1}\mathbf{X}_c^T\mathbf{y} \qquad (2.9)$$

$$b_0 = \bar{y} - \bar{\mathbf{x}}^T\mathbf{b}_1 \qquad (2.10)$$

where $\mathbf{b} =[b_0; \mathbf{b}_1]$, $\mathbf{y}_c$ is the column-centered vector **y**, $\bar{y}$ is the mean value of the elements of **y** and $\bar{\mathbf{x}}$ is the vector of the means of the columns of $\mathbf{X}_{-1}$. Centered data is frequently used in multivariate calibration where the preferred notation for the model equation is (see § 2.4.2):

$$\mathbf{y} = \mathbf{1}\beta_0 + \mathbf{X}\beta + \varepsilon \qquad (2.11)$$

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

In this case, **X** corresponds to the matrix $\mathbf{X}_{-1}$ defined above. It must always be clear if either eq 2.6 or eq. 2.11 is being used. Below, the **X** of eq 2.6 is used.

The variance-covariance matrix of **b** is:

$$\text{var}(\mathbf{b}) = (\mathbf{X}^T\mathbf{X})^{-1}\,\sigma^2 = \begin{pmatrix} \text{var}(b_0) & \text{cov}(b_0,b_1) & \cdots & \text{cov}(b_0,b_J) \\ \text{cov}(b_0,b_1) & \text{var}(b_1) & \cdots & \text{cov}(b_1,b_J) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(b_J,b_0) & \text{cov}(b_J,b_1) & \cdots & \text{var}(b_J) \end{pmatrix} \quad (2.12)$$

The variance of the coefficient b_j is

$$var(b_j) = \text{UVIF}_j\,\sigma^2 \quad (2.13)$$

where the *variance coefficient*[10] UVIF_j is the jth element in the diagonal of $(\mathbf{X}^T\mathbf{X})^{-1}$. The off-diagonal elements of $(\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$ are the covariances between the coefficients. The expression for column-centered data is:

$$\text{var}(\mathbf{b}_1) = (\mathbf{X}_c^T\mathbf{X}_c)^{-1}\,\sigma^2 \quad (2.14)$$

The boundary of the 100(1-α) per cent confidence region for all the coefficients is[11,14-16]:

$$(\beta - \mathbf{b})^T\,\mathbf{X}^T\mathbf{X}\,(\beta - \mathbf{b}) = p^2 \quad (2.15)$$

where $p^2 = Q\,s^2\,F_{Q,df,\alpha}$, s^2 is an estimate of σ^2 with df degrees of freedom and $F_{Q,df,\alpha}$ is the α per cent point of the F distribution with Q and df degrees of freedom. In the Q-dimensional space of the coefficients, eq 2.15 defines an hyperellipsoid centered in **b** (an ellipse for designs with two coefficients). The true values of the coefficients are supposed to be inside that hyperellipsoid with a probability α to commit a type I error. A plot of the confidence region for the case of two and three coefficients is shown in the Figure 2.3. Many characteristics of the ellipsoid are related to the

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

eigenvalues (λ_a) and eigenvectors (v_a) of $(X^TX)^{-1}$. The ath axis is oriented in the direction of v_a and its half-length is $p\sqrt{\lambda_a}$. In addition, $\text{Det}(X^TX)^{-1} = \prod_a \lambda_a$ and $\text{Tr}(X^TX)^{-1} = \sum_a \lambda_a$.

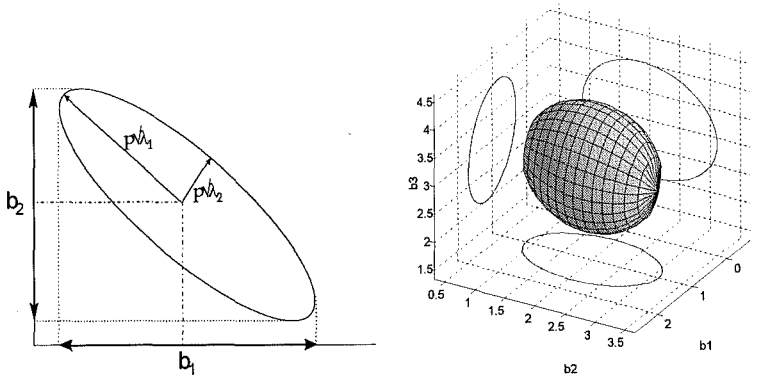


Figure 2.3. Confidence ellipses of a two-coefficient model (left) and a three-coefficient model (right).

2.3.3.3 Prediction

The predicted response value for the input variables $x_{un} = [x_{un,1}, x_{un,2}, \ldots x_{un,J}]^T$ is:

$$y_{un} = \mathbf{f}^T(x_{un})\,\mathbf{b} \qquad (2.16)$$

and has an associated variance (due to the propagation of the uncertainty of the coefficients):

$$var(y_{un}) = \mathbf{f}^T(x_{un})\,(X^TX)^{-1}\,\mathbf{f}\,(x_{un})\,\sigma^2 = d(x_{un})\,\sigma^2 \qquad (2.17)$$

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

where $d(\mathbf{x}_{un})= \mathbf{f}^T(\mathbf{x}_{un}) (\mathbf{X}^T\mathbf{X})^{-1} \mathbf{f}(\mathbf{x}_{un})$ is called the *variance function*. For the model in eq 2.11, the prediction is:

$$y_{un}= b_0 + \mathbf{f}^T(\mathbf{x}_{un})\, \mathbf{b} \qquad (2.18)$$

For the particular case of a first degree polynomium :

$$y_{un} = b_0 + b_1 x_{un,1} + b_2 x_{un,2} + \ldots + b_J x_{un,J} = b_0 + \mathbf{x}_{un}{}^T \mathbf{b} \qquad (2.19)$$

2.3.3.4 Geometrical interpretation of the least squares solution

The geometrical interpretation of the least-squares solution is the underlying basis for the calculation of the net analyte signal[17] and of the quantification using the classical least squares (CLS) model (see § 2.4.1.1). Each column of $\mathbf{X}$ can be regarded as a vector in a Euclidean space so that all the columns define an (hyper)plane Π. The vectors $\eta=\mathbf{X}\beta$ and $\mathbf{y}_{pred}=\mathbf{X}\mathbf{b}$ are linear combinations of the columns of $\mathbf{X}$ and lie in the (hyper)plane. However the vector $\mathbf{y}=\mathbf{X}\beta+\varepsilon$ does not lie in this (hyper)plane since ε may be different from zero. The least-squares estimation of β is the vector $\mathbf{b}$ that gives a residual vector $\mathbf{e}=\mathbf{y}-\mathbf{y}_{pred}$ of minimal length (i.e. $\mathbf{e}^T\mathbf{e}$ = minimal, which is the sum of the squared errors). This happens when $\mathbf{y}_{pred}$ is the orthogonal projection of $\mathbf{y}$ onto the (hyper)plane and then $\mathbf{e}$ is orthogonal to this (hyper)plane. Since $\mathbf{b}= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, then $\mathbf{y}_{pred} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{H}\mathbf{y}$ and $\mathbf{e}=(\mathbf{I}-\mathbf{H})\mathbf{y}$ where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is a *projection matrix*. Figure 2.4 represents this situation in a three-dimensional space.

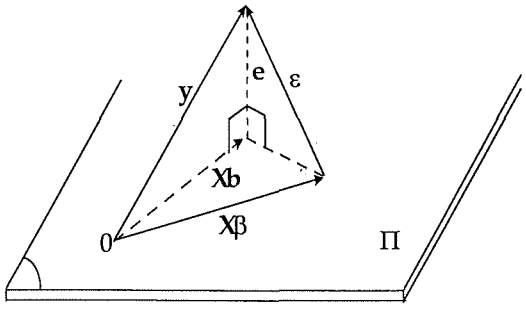


Figure 2.4. Geometrical interpretation of the least-squares solution.

2.3.3.5 Interpretation of the estimated coefficients

The coefficients of the models considered give the rate of change of the dependent variable for a unit change in the independent variable when all of the other variables are held constant. By properly scaling the independent variables (see §3.2), the coefficients enable the effect in the response of each independent variable (main effect) or combination of variables (interaction effects) to be compared[18]. A treatment is judged statistically significant if the variation in the response caused by changing the variable setting (or combination of variable settings) is larger than the experimental error in the measurement of the response. Otherwise, the variable is judged as statistically insignificant. This approach is used in experimental plans based on matrices such as Hadamard, factorial, fractional factorial, etc. [19,20].

2.3.4 Optimal design in multiple linear regression

Eqs 2.12 and 2.17 show that the precision of the estimated coefficients and of the predicted response depends on the matrix **X** but not on the values of y. The important consequence is that, before carrying out any experiment, the settings of the independent variables in **X** (called the *design of experiments*) can be decided to achieve the quality of the properties demanded to the model (e.g. the best global precision of **b** or an acceptable precision of y_{un}). The cost of the model can be reduced by finding the design with the minimum number of experiments that contains the required information (see Scheme 2.1 at the end of §2.3). Several criteria (see §2.3.4.2) enable to evaluate the quality of a design using only the **X** matrix.

Experimental designs are routinely used in many areas of scientific investigation. Catalogs of the most appropriate designs linked to different types of regression models are available [5,16,21]. Some examples are the factorial, Hadamard and Doehlert designs. The importance of this kind of designs (called *classical designs*) lies in two points: they give the maximum information with a small number of experiments by varying several variables simultaneously; and they have the ability of uncorrelating the estimated coefficients (e.g. in factorial and fractional factorial

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

designs the experiments are chosen so that the columns of **X** are orthogonal to each other). One of their limitations is that the variable settings must be independently adjusted in the same experiment according to the design. This can not always be accomplished. Sometimes the variables are correlated by necessity or some experiments cannot be performed due to an irregularly shaped experimental domain. In addition, models that deviate from the usual first or second order ones may have no classical design. Moreover the number of experiments in the classical design may be too large for the experimenter's purposes. However, using non-designed experiments has the danger of leading to highly collinear columns in **X** and even situations where $\mathbf{X}^T\mathbf{X}$ is not of full rank.

When the requirements of classical experimental designs cannot be met, the permitted design space can be examined to search for the optimal experiments that achieves the necessary quality of the properties of the model (e.g. precise estimates of the coefficients of the model). This requires:

1. A matrix of N candidate points (experiments) (ξ_N) where each row represents one experiment and each column a variable. Sets of I candidate points (ξ_I) are chosen from this matrix. The set of all possible ξ_I is designed by Ξ.

2. A criterion to quantify the quality of a proposed set ξ_I. The optimal set optimizes this criterion among all other possible sets of the same number of experiments.

3. An algorithm for finding the optimal set and avoid checking all possible combinations of experiments in Ξ.

4. A criterion to decide the optimal number of experiments I.

These requirements are general for optimization problems of combinatorial nature and are considered in the next sections applied to the DOE. They are valid for sample and wavelength selection considered in the chapters 3 and 4.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

2.3.4.1 List of candidate points

The quality of all the candidate points is of primary concern for the quality of the optimal set selected from this list. If the candidate points do not meet the requirements to built the model, the selected points may not be suitable despite being optimal for the given list. In multivariate calibration this implies that the preparation or selection of calibration samples must span all known sources of variation present in analysis samples, and the quality of the selected sets must be checked to assess that the sensors (or samples) selected as optimal are good enough for the experimenter's purposes. Some algorithms can search the design space and do not need a list of candidate points[22].

2.3.4.2 Optimality criteria for experimental designs

To base optimization decisions on, the *goodness* of each design is evaluated by a mathematical function called a *criterion function*. Criterion functions for optimal parameter estimation, model discrimination, robust regression, design augmentation, etc. have been classified and explained in MLR [23-28]. However, many publications usually employ mathematical and statistical language that requires a large effort for a chemist to understand it. Some easily readable texts are references: 9, 14, 15, 16, 21, 29, 30, 31 and the review by Steinberg[32]. Among the many criteria available, the most popular consider the quality of the estimated coefficients (which are related to the volume, shape, and orientation of the confidence ellipsoid (Figure 2.3)) and to the variance of the predicted response[32,33] . These optimality criteria are the following *:

1. *D-criterion.* A design is *D-optimal* if it minimizes $\mathrm{Det}(\mathbf{X}^T\mathbf{X})^{-1}$ over all other possible designs in Ξ. Since $\mathrm{Det}(\mathbf{X}^T\mathbf{X})=1/\mathrm{Det}\ (\mathbf{X}^T\mathbf{X})^{-1}$, a D-optimal design also maximizes $\mathrm{Det}(\mathbf{X}^T\mathbf{X})$. It is said that this design minimizes the generalized variance of the least squares estimate of β (low values for the variances and covariances of the coefficients). The volume of the confidence ellipsoid is

* The value of σ^2 ,assumed constant during the experimentation, is not considered in the comparison of experimental designs since it is the same for all of them.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

proportional to the product of the length of the half-axes [14-16,26,27,34], thus to $\Pi\sqrt{\lambda_a}=[\mathrm{Det}(\mathbf{X}^T\mathbf{X})]^{-1/2}$. A D-optimal design minimizes the volume of the confidence region: the larger the determinant, the smaller the volume and the larger the precision of the coefficients. D-optimality is a very popular criterion. Its drawback is that the volume of the ellipsoid of a D-optimal design can be small because it is "narrow but long"[26]. This corresponds to a list of candidate samples that are quite collinear. Therefore the optimal subset should be always checked for its quality.

2. *A-criterion.* A design is *A-optimal* if it minimizes $\mathrm{Tr}(\mathbf{X}^T\mathbf{X})^{-1}$ over all other possible designs in Ξ . This is related to the shape of the ellipsoid and to the average variance of the coefficients. The ellipsoid of the A-optimal set is the smallest hypersphere of radius r given by $r^2=\mathrm{Tr}[\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}]$.

3. *E-criterion.* A design is *E-optimal* if it minimizes the maximal eigenvalue of $(\mathbf{X}^T\mathbf{X})^{-1}$ over all other possible designs in Ξ, thus the largest principal axis of the uncertainty ellipsoid has minimum length. This is related to the shape of the ellipsoid. The ellipsoid of the E-optimal set is the smallest hypersphere of radius r given by $r^2=$ maximum eigenvalue of $[\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}]$.

4. *G-criterion.* A design is G-optimal if it minimizes the maximum value of the variance of the predicted response at $\mathbf{x}_{un}$ $var(y_{un})$ given by eq 2.17 over all possible $\mathbf{x}$ in the design space. D- and G-optimal designs are equivalent in continuous designs [32] (designs were the distribution of experiments is given as a measure[15,27]). However, this does not always hold for exact designs (designs for a specified number of experiments I) where the D- and G-optimal may be not the same[15,29].

5. *Condition number of* $(\mathbf{X}^T\mathbf{X})^{-1}$. It is defined as $\mathrm{Cond}(\mathbf{X}^T\mathbf{X})^{-1}=\max(\lambda_a)/\min(\lambda_a)$. This a measure of eccentricity of the ellipsoid, and does not depend either on the volume or on the orientation of the ellipsoid.

6. *Variance inflation factors.* This criterion is related to the orientation of the ellipsoid. It is considered in §2.6.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

2.3.4.3 Algorithms for experiment selection

Searching the best subset in a design space made of a large list of candidate experiments is a combinatorial problem that may be time-prohibitive if all possible combinations must be checked. Optimization algorithms can find the optimum (or an acceptable solution) in a reasonable time. Among the algorithms specialized for optimizing a specific criterion[26], the ones for D-optimal designs are very popular [15, 31, 32] ,specially Fedorov's exchange algorithm[16,31,35]. Fedorov's algorithm uses the list of candidate points expressed as a model matrix. Among the general optimization procedures, genetic algorithms (GA) and generalized simulated annealing (GSA) [36] have been used to find optimal designs. GA has advantages over Fedorov's algorithm in the search of D-optimal matrices with many experimental points and a large list of candidates[22]: less computation time and no need of a list of candidate points. These algorithms compete with others based on spanning the experimental design as much as possible such as clustering[37] or the Kennard-Stone algorithm[38,39]. The Fedorov's algorithm is considered in §3.2.

2.3.4.4 Criterion to decide the optimal number of experiments

The optimal number of experiments depends on the constraints imposed by the cost of the experiments, the time and the difficulty to perform them, the precision required, etc. Only the criterion derived from the D-optimality is considered here. $\mathrm{Det}(\mathbf{X}^T\mathbf{X})$ measures the information of the design and always increases when a new experiment (a new row) is added to $\mathbf{X}$[22]. Thus, the more experiments are performed, the more information the design has and the precision of the coefficients increases. However $\mathrm{Det}(\mathbf{X}^T\mathbf{X})$ is not useful for comparing designs with a different number of experiments: a relatively small increase in the precision may not justify carrying out an extra experiment. Instead, the *normalized information matrix*, (or *matrix of moments*[35]) $\mathbf{M}=(\mathbf{X}^T\mathbf{X})/I$ is used. Then $\mathrm{Det}(\mathbf{M})=\mathrm{Det}(\mathbf{X}^T\mathbf{X})/I^Q$, where Q is the number of coefficients in the model, is independent of the number of experiments and gives a measure of the amount of information *per experiment*. This enables designs with a different number of points to be compared. The design with maximum $\mathrm{Det}(\mathbf{M})$ is selected as optimal. However, M-optimality has selected

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

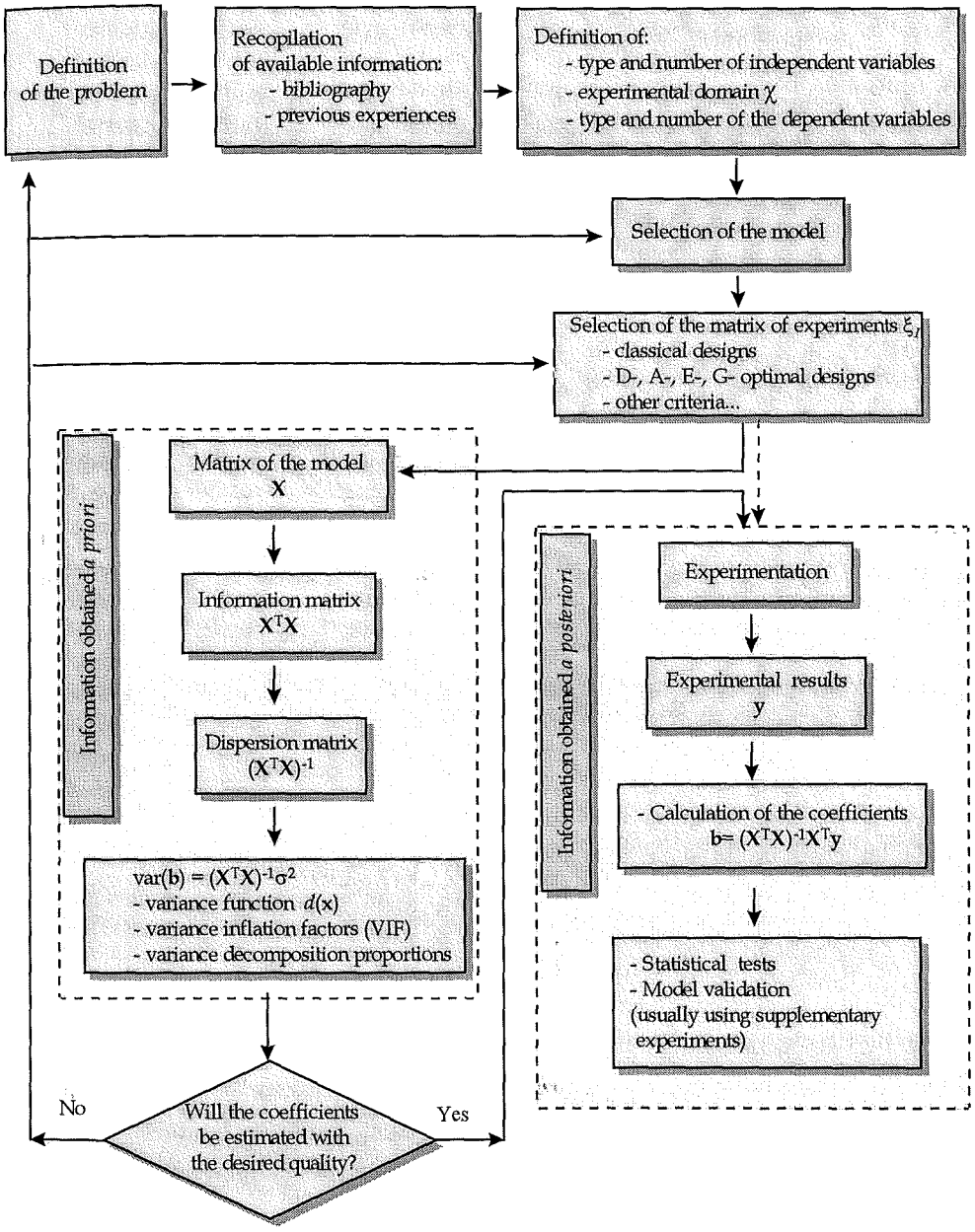
the best points from the list of candidates. If the list is made of highly collinear points, the M-optimal solution can be unacceptable. For this reason, the variance inflation factors (VIFs) should be calculated for the M-optimal subset to assess that it can be used for regression.

2.3.4.5 Objections to some experimental design criteria

The main objection to using the cited optimal experimental designs is their dependence on a model that must be postulated completely in advance[37,40]. Consequently, the selected points can be unsuitable for another model and provide no means for testing the structure of the model or to estimate possible additional effects. Zemroch[40] discussed model sensitivity of an experimental design and said that single- and multifactor designs with evenly spread points over the actual region will not have this drawback. However, optimal designs have some excellent statistical properties and enable to obtain the maximum information with the minimum number of experimental runs. They are the best option if some *a priori* knowledge of the model is assumed.

2.3.4.6 Summary of optimal experimental design

Scheme 2.1 summarizes the ideas of the methodology for optimal experimental design: using the necessary references (bibliography, needs of the experimenter, purposes of the experimentation, etc.) the experimental domain of interest χ is decided and a mathematical model is postulated. Then, the matrix of experiments ξ_I is selected, either using catalogs of designs or algorithms that optimize the appropriate selection criterion. *A priori* criteria of quality are evaluated from the model matrix **X**. If the matrix contains the necessary information to provide a model with the desired qualities the experiments are performed and the coefficients of the model calculated. Finally, statistical tests are made on the coefficients and/or the model is validated usually with supplementary experiments. The information on the left hand column in the scheme is known before experimentation and it helps to choose the best experimental conditions.



Scheme 2.1 Summary of the procedure for optimal experimental design.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

2.4.Multivariate calibration models

Multivariate calibration models are used in chemical analysis for predicting the concentration of the analyte k in an unknown sample ($c_{un,k}$) from the vector of measured responses of this sample at J sensors $\mathbf{r}_{un}^T=[r_{un,1}, ..., r_{un,J}]$. The most common prediction model is the linear equation (Figure 2.5):

$$c_{un,k} = b_{0,k} + r_{un,1}b_{1,k} + ... + r_{un,J}b_{J,k} = b_{0,k} + \mathbf{r}_{un}^T \mathbf{b}_k \qquad (2.20)$$

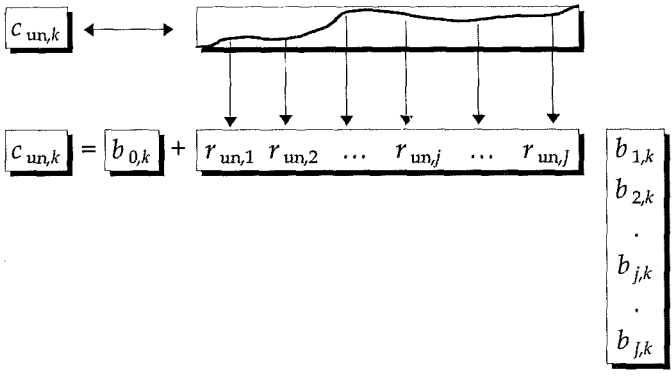


Figure 2.5. Prediction of the concentration in an unknown sample from its digitalized instrumental responses.

where $\mathbf{b}_k=[b_{1,k}, ... , b_{J,k}]^T$. Using the responses and known analyte concentrations of the *calibration samples*, the coefficients $b_{j,k}$ are estimated in a way that distinguishes the different regression methods[41-53]. The multivariate calibration techniques considered in this thesis are: classical least squares (CLS), inverse least squares (ILS), principal component regression (PCR) and partial least squares (PLS). They are also called first-order calibration models since a response vector is used for each sample[42,43,52]. Their underlying theory, advantages and limitations are explained in this section.

2.4.1 Direct models

2.4.1.1 Classical least squares (CLS)

This quantification method, also known as K-matrix calibration[47,51,54-56], is based on the Beer's law applied to many wavelengths. The model assumes that the measured absorbance at each wavelength is a linear additive function of the concentrations of the chemical constituents and that there are no interaction terms in the spectrum between the various components of the sample. The model equation for a sample i of K analytes is:

$$r_{i,j} = r_{0,j} + \sum_{k=1}^{K} s_{j,k}\, c_{i,k} + e_{i,j} \qquad (2.21)$$

where $r_{i,j}$ is the response measured at the jth sensor, $s_{j,k}$ is the sensitivity (the response divided by the concentration of the analyte in a pure sample), $r_{0,j}$ is the background contribution, $c_{i,k}$ is the concentration of the analyte k in the mixture and $e_{i,j}$ is the measurement error that is supposed independent and normally distributed $e_{i,j} \sim N(0,\sigma^2)$. The model error is assumed to derive from the measurement of the absorbances. The model for J measured wavelengths after accounting for the background (for example by substracting from each measured spectrum the response of a blank containing only the sample matrix, without the analytes of interest) is (Figure 2.6) :

$$\mathbf{r} = \mathbf{S}\, \mathbf{c}_{true} + \boldsymbol{\varepsilon} \qquad (2.22)$$

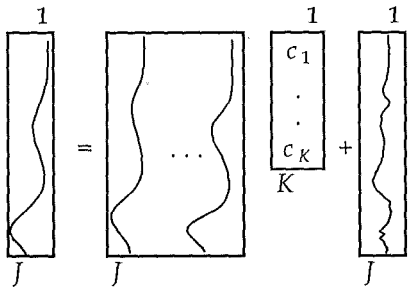


Figure 2.6 Matrix representation of the CLS model.

where $\mathbf{r}_{J\times1}$ is the response vector of the sample, $\mathbf{c}_{true}$ is the $K\times1$ vector of all analyte concentrations in the sample that give response, $\mathbf{S}_{J\times K}=[\mathbf{s}_1, ..., \mathbf{s}_k, ..., \mathbf{s}_K]$ is a matrix of sensitivities whose columns are the pure-component spectra at unit concentration and unit path length (absorptivity-path length products[51]) and $\varepsilon_{J\times1}$ is a vector of errors.

2.4.1.1.1 Calibration

Calibration consists of calculating **S**. Eq 2.22 for I calibration samples is:

$$\mathbf{R}^T=\mathbf{S}\mathbf{C}^T + \mathbf{E} \tag{2.23}$$

and **S** is estimated as:

$$\mathbf{S} = (\mathbf{C}^+\mathbf{R})^T \tag{2.24}$$

where $\mathbf{R}_{I\times J}$ contains the measured absorbances at J wavelengths for I calibration solutions of individual components or mixtures, $\mathbf{C}_{I\times K}$ are the known concentrations of the K analytes in the calibration samples and $\mathbf{E}_{I\times J}$ is the matrix of spectral errors. If the number of calibration samples and constituents is the same **C** is square ($K\times K$) and **S** can be estimated as:

$$\mathbf{S} = (\mathbf{C}^{-1}\mathbf{R})^T \tag{2.25}$$

In addition, if each calibration sample contains only one analyte, $\mathbf{C}_{K\times K}$ is diagonal and eq 2.25 just calculates **S** by dividing the spectrum of each calibration sample by its analyte concentration. To obtain a better estimation of **S** the number of calibration samples is usually larger than the number of components. To calculate **S** in the above equations, the relative amounts of constituents in at least K calibration samples must change from one sample to another. That means that only dilutions of one concentrated calibration sample cannot be used (see § 2.6.5 for the collinearity problem). In addition, to have the necessary linearly independent equations in the prediction step, the number of wavelengths must be equal or larger than the number of constituents in the mixtures ($J\geq K$). Usually the entire spectrum is used.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

2.4.1.1.2 Prediction

The least-squares estimation of the concentrations of the K analytes in the unknown sample is:

$$\mathbf{c}_{un} = \mathbf{S}^{+}\mathbf{r}_{un} \tag{2.26}$$

In ordinary least squares $\mathbf{S}^{+}$ is calculated as $(\mathbf{S}^{T}\mathbf{S})^{-1}\mathbf{S}^{T}$. The concentration of the analyte k (the kth element of $\mathbf{c}_{un}$) is obtained by multiplying the kth row of $\mathbf{S}^{+}$ by $\mathbf{r}_{un}$:

$$c_{un,k} = \mathbf{S}^{+}_{k\text{-row}}\,\mathbf{r}_{un} = \mathbf{r}_{un}^{T}\,\mathbf{S}^{T+}_{k\text{-col}} \tag{2.27}$$

since $(\mathbf{S}^{+})^{T} = \mathbf{S}^{T+}$. The expressions for the variance of the predicted concentrations and related measures can be found in §4.3. Comparing the last term in eq 2.27 with eq 2.20, it can be seen that, considering $b_{0,k}=0$, the kth-column of $\mathbf{S}^{T+}$ is the vector of coefficients:

$$\mathbf{b}_{k,CLS} = \mathbf{S}^{T+}_{k\text{-col}} \tag{2.28}$$

(also $\mathbf{B}_{CLS}=[\mathbf{b}_{1,CLS}, \mathbf{b}_{2,CLS}, ..., \mathbf{b}_{K,CLS}]=\mathbf{S}^{T+}$ and $\mathbf{B}_{CLS}^{T}\mathbf{B}_{CLS} = (\mathbf{S}^{T}\mathbf{S})^{-1}$). This shows that, to quantify, it is not necessary to know $\mathbf{S}$ but only $\mathbf{S}^{T+}$ that can be estimated as[1,57]:

$$\mathbf{S}^{T+} = (\mathbf{C}^{+}\mathbf{R})^{+} = \mathbf{R}^{+}\mathbf{C} \tag{2.29}$$

and the vector of coefficients for the analyte k is:

$$\mathbf{b}_{k,ILS} = \mathbf{R}^{+}\mathbf{c}_{k} \tag{2.30}$$

where $\mathbf{c}_k$ is the column in $\mathbf{C}$ of the concentrations of the analyte k in all the calibration samples. These coefficients correspond to the P-matrix calibration (§2.4.2.1). Eqs 2.28 and 2.30 have been used to justify that the K-matrix and P-matrix approaches are basically the same model[1], and that the K-matrix calibration is a particular case of the P-matrix calibration where the concentration of all the analytes that give response are simultaneously predicted. However, the equality $\mathbf{b}_{k,CLS} = \mathbf{b}_{k,ILS}$ is only true when $\mathbf{R}^{T}=\mathbf{S}\mathbf{C}^{T}$, that is to say, when there is no error term

in eq 2.23. Since errors are always present because **R** is made of measured quantities $\mathbf{b}_{k,CLS}$ and $\mathbf{b}_{k,ILS}$ are normally slightly different.

Prediction using the net analyte signal

The *net analyte signal* (NAS)[17], is the part of the spectrum of a component in a mixture that is orthogonal to the spectra of the other components in that mixture. In CLS the net signal of the analyte k in the pure component spectra and in the spectrum of the unknown sample are given by:

$$\mathbf{s}_k^* = (\mathbf{I} - \mathbf{S}_k\mathbf{S}_k^+)\mathbf{s}_k \quad (2.31)$$

$$\mathbf{r}_{un,k}^* = (\mathbf{I} - \mathbf{S}_k\mathbf{S}_k^+)\mathbf{r}_{un} \quad (2.32)$$

where $\mathbf{S}_k = [\mathbf{s}_1, \ldots, \mathbf{s}_{k-1}, \mathbf{s}_{k+1}, \ldots \mathbf{s}_K]$ is the **S** matrix without the kth column. $\mathbf{s}_k^*$ is the vector of residuals of the regression of $\mathbf{s}_k$ versus $\mathbf{S}_k$ ($\mathbf{s}_k = \mathbf{S}_k\mathbf{b}^* + \mathbf{s}_k^*$) since:

$$\mathbf{s}_k^* = (\mathbf{I} - \mathbf{S}_k\mathbf{S}_k^+)\mathbf{s}_k = \mathbf{s}_k - \mathbf{S}_k\mathbf{S}_k^+\mathbf{s}_k = \mathbf{s}_k - \mathbf{S}_k\mathbf{b}^* = \mathbf{s}_k - \mathbf{s}_{k,proj} = \text{residuals}$$

This is related to the geometrical interpretation of the least-square solution (see §2.3.3.4). The spectra of the other pure components ($\mathbf{S}_k$) define a (hyper)plane (Figure 2.7) so that $\mathbf{s}_{k,proj} = \mathbf{S}_k\mathbf{b}^*$ is the part of $\mathbf{s}_k$ that lies in that (hyper)plane because it is a linear combination of the columns of $\mathbf{S}_k$. The remaining part of the signal, the residuals $\mathbf{s}_k - \mathbf{s}_{k,proj}$, are orthogonal to the (hyper)plane and can be used for quantification. $\mathbf{S}_k\mathbf{S}_k^+$ is a *projection matrix*[1] of the vector onto the space spanned by $\mathbf{S}_k$ and $(\mathbf{I} - \mathbf{S}_k\mathbf{S}_k^+)$ is projection matrix onto a space orthogonal to that spanned by $\mathbf{S}_k$. The same stands for $\mathbf{r}_{un,k}^*$. The vector $\mathbf{s}_k^* / \|\mathbf{s}_k^*\|^2$ contains the regression coefficients and is different for each analyte but the same for all the unknown samples. The $\mathbf{s}_k^*$ of the different analytes are not necessarily orthogonal among them. $\mathbf{r}_{un,k}^*$ is different for each analyte and unknown sample. In absence of the

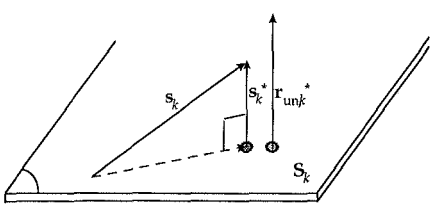


Figure 2.7. The net analyte signal of the analyte k and of the unknown sample spectrum for a model without error.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

error term in eq 2.22 the vectors $\mathbf{r}_{un,k}^*$ and $\mathbf{s}_k^*$ are parallel and the proportionality constant is the concentration of the analyte k (see §4.7):

$$\mathbf{r}_{un,k}^* = c_{un,k,true}\,\mathbf{s}_k^* \tag{2.33}$$

so that $c_{un,k,true}$ can be calculated using either the net analyte signal at any wavelength j or the norm of the net analyte signals:

$$c_{un,k,true} = \mathbf{r}_{un,j,k}^* / \mathbf{s}_{j,k}^* = \| \mathbf{r}_k^* \| / \| \mathbf{s}_k^* \| \tag{2.34}$$

The vectors in eq 2.33 are represented in Figure 2.7. Since the error term in eq 2.22 is present the relationship is:

$$\mathbf{r}_{un,k}^* = c_{un,k,true}\,\mathbf{s}_k^* + \boldsymbol{\varepsilon}_k^* \tag{2.35}$$

where $\boldsymbol{\varepsilon}_k^*=(\mathbf{I}-\mathbf{S}_k\mathbf{S}_k^+)\boldsymbol{\varepsilon}$ is the net analyte signal of the error, which is unknown. Then $c_{un,k}$, the least squares estimation of $c_{un,k,true}$, is found as:

$$\begin{aligned}
&\mathbf{r}_{un,k}^* = c_{un,k}\,\mathbf{s}_k^* + \mathbf{e}_k^* \\
&\mathbf{r}_{un,k}^{*T} = c_{un,k}\,\mathbf{s}_k^{*T} + \mathbf{e}_k^{*T} \\
&\mathbf{r}_{un,k}^{*T}\mathbf{s}_k^* = c_{un,k}\,\mathbf{s}_k^{*T}\mathbf{s}_k^* + \mathbf{e}_k^{*T}\mathbf{s}_k^* \\
&c_{un,k} = \mathbf{r}_{un,k}^{*T}\mathbf{s}_k^* / \| \mathbf{s}_k^* \|^2
\end{aligned}$$

where $\mathbf{e}_k^*$ is the part of $\mathbf{r}_{un,k}^*$ not explained by $c_{un,k}\,\mathbf{s}_k^*$ and orthogonal to $\mathbf{s}_k^*$ and $\mathbf{e}_k^{*T}\mathbf{s}_k^*=0$. Using the idempotency property of the projection matrices:

$$\mathbf{r}_{un,k}^{*T}\mathbf{s}_k^* = \mathbf{r}_{un}^T(\mathbf{I}-\mathbf{S}_k\mathbf{S}_k^+)^T(\mathbf{I}-\mathbf{S}_k\mathbf{S}_k^+)\mathbf{s}_k = \mathbf{r}_{un}^T(\mathbf{I}-\mathbf{S}_k\mathbf{S}_k^+)\mathbf{s}_k = \mathbf{r}_{un}^T\mathbf{s}_k^* \tag{2.36}$$

so that the concentration of the analyte k in the unknown sample is given by:

$$c_{un,k} = \mathbf{r}_{un}^T\mathbf{s}_k^* / \| \mathbf{s}_k^* \|^2 \tag{2.37}$$

Comparing eq 2.20, 2.28 and 2.35, the vector of regression coefficients of the CLS model for the analyte k is :

$$\mathbf{b}_{k,CLS} = \mathbf{S}^{T+}_{k\text{-col}} = [\mathbf{S}^{+}_{k\text{-row}}]^T = \mathbf{s}_k^* / \|\mathbf{s}_k^*\|^2 \quad (2.38)$$

So that the prediction equation for CLS is:

$$c_{un,k} = \mathbf{S}^{+}_{k\text{-row}}\, \mathbf{r}_{un} = \mathbf{b}_{k,CLS}^T\, \mathbf{r}_{un} = \mathbf{r}_{un}^T \mathbf{b}_{k,CLS} \quad (2.39)$$

The mathematical expressions of the variance of the predicted concentrations, the selectivity and sensitivity in CLS can be found in §4.2.3. The following equalities can be deduced from the equations above [59]:

$$\|\mathbf{b}_{k,CLS}\|^2 = (\mathbf{S}^T\mathbf{S})^{-1}_{kk} = \|\mathbf{S}^{+}_{k\text{-row}}\|^2 = 1/\|\mathbf{s}_k^*\|^2 \quad (2.40)$$

$$\mathbf{b}_{k,CLS} / \|\mathbf{b}_{k,CLS}\| = (\mathbf{s}_k^*/\|\mathbf{s}_k^*\|^2)/(1/\|\mathbf{s}_k^*\|) = \mathbf{s}_k^*/\|\mathbf{s}_k^*\| \quad (2.41)$$

$$\mathbf{b}_{k,CLS}^T \mathbf{s}_k^* = 1 \quad (2.42)$$

where $(\mathbf{S}^T\mathbf{S})^{-1}_{kk}$ is the kth diagonal element of $(\mathbf{S}^T\mathbf{S})^{-1}$.

The expression in CLS can also be formulated[17] using $\mathbf{a}_k$, the spectrum of the analyte k at concentration c_k^0. This vector is related to the sensitivities for k analyte as: $\mathbf{a}_k = c_k^0 \mathbf{s}_k$ and therefore the norm of the vectors follows $\|\mathbf{a}_k\| = c_k^0 \|\mathbf{s}_k\|$. The NAS of $\mathbf{a}_k$ is $\mathbf{a}_k^* = (\mathbf{I} - \mathbf{A}_k\mathbf{A}_k^+)\mathbf{a}_k$ where $\mathbf{A}_k$ is the matrix $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_k \ldots, \mathbf{a}_K]$ with the ath column deleted ($\mathbf{A} = \mathbf{S}\mathbf{C}_0$, $\mathbf{S} = \mathbf{A}\mathbf{C}_0^{-1}$ and $\mathbf{S}^+ = \mathbf{C}_0\mathbf{A}^+$ where $\mathbf{C}_0$ is a $K \times K$ diagonal matrix whose diagonal elements are c_k^0). The NAS is related to $\mathbf{s}_k^*$ with $\mathbf{a}_k^* = c_k^0\, \mathbf{s}_k^*$ (and also $\mathbf{a}_k^*/\|\mathbf{a}_k^*\| = \mathbf{s}_k^*/\|\mathbf{s}_k^*\|$). The net analyte signal of the unknown sample is calculated as $\mathbf{r}_{un,k}^* = (\mathbf{I} - \mathbf{A}_k\mathbf{A}_k^+)\mathbf{r}_{un} = (\mathbf{I} - \mathbf{S}_k\mathbf{S}_k^+)\mathbf{r}_{un}$ and it is related to the concentration as: $\mathbf{r}_{un,k}^* = c_{un,k}\, \mathbf{s}_k^* = c_{un,k}\, \mathbf{a}_k^*/c_k^0$. Therefore, the predicted concentration in the unknown sample can be calculated as $c_{un,k} = c_k^0\, \mathbf{r}_{un,k}^{*T} \mathbf{a}_k^*/\|\mathbf{a}_k^*\|^2 = c_k^0\, \mathbf{r}_{un,k}^T \mathbf{a}_k^*/\|\mathbf{a}_k^*\|^2$

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

2.4.1.1.3 Advantages of CLS

1. Unlike the univariate calibration, CLS does not require selective measurements.

2. The Beer's Law provides a sound foundation for the predictive model.

3. This model can be used for mixtures of known qualitative composition. (e.g., gas phase spectroscopy, some process monitoring or pharmaceutical samples).

4. The model can use a large number of wavelengths to gain a signal averaging effect[43] beneficial for the precision of the predicted concentration, making it less susceptible to noise in the spectra.

5. It may provide a reasonable basis for extrapolation and understanding of the uncertainty in predicted values of the analyte.

2.4.1.1.4 Limitations of CLS [51,60]

1. CLS can only be applied to systems where the spectrum of all the pure constituents giving rise to a signal is known since equations assume the response at a wavelength is due entirely to the calibrated constituents. If the response of the unknown sample contains signal from constituents that have not been included in the calibration matrix **S** in addition to background problems and baseline effects, biased predicted concentrations can be obtained. A CLS model built using second-derivative spectra has been reported to solve partially this problem[61].

2. CLS is not useful for mixtures with interaction between constituents or deviations from Beer's law (nonlinear calibration curves).

3. A severe overlap of spectral bands, quite usual in ultraviolet and visible (UV-vis) spectra, can introduce large uncertainty in the estimated concentrations. This is considered more extensively in the section §2.6.

2.4.2 Inverse Models

The limitations of CLS would make this method fail in the analysis of many common samples such as natural products (water, flour, meat,...) whose complex compositional chemistry makes it impossible to know the spectra of all the pure components that give response. Moreover, sometimes only the quantities of not all but only some of the constituents are of interest. The spectroscopic quantitative analysis of samples with complex matrices can be made using the *inverse* calibration:

$$c_k = f(r_1, r_2, \ldots r_J) + \text{residual}\, f_k$$

where the concentration of the analyte of interest, c_k, is modeled as a function of the instrumental measurements $r_1, r_2, \ldots r_J$ (e.g. absorbances at selected wavelengths) following a empirical relationship, without a theoretical underlying such as the Beer's law. These methods, which include ILS, PCR or PLS, can build calibration models without knowledge of the concentrations of all the constituents in the calibration set. The concentration of only the analyte of interest in each calibration sample (or matrix of concentration of the constituents of interest, C) is needed. Thus, so that unknown interferents can be present in the calibration samples. This important advantage, common for ILS, PCR and PLS, will not be mentioned again. These models have been applied successfully to both natural and manufactured products such as wheat, meat, gasoline or plastics. Although samples with known interfering species do not need to be prepared, the samples must contain the analytes and interferences which contribute to the response so that all the causes of variation can be considered in the model.

Calibration of inverse models

In the ILS, PCR and PLS models considered here the concentration of the analyte of interest k in the sample i is regressed as a linear combination of the instrumental measurements at J selected sensors[3,44,46,62]:

$$c_{i,k} = \beta_{0,k} + \beta_{1,k} r_{i,1} + \ldots + \beta_{J,k} r_{i,J} + \varepsilon_{i,k} = \beta_{0,k} + \mathbf{r}_i^T \beta_k + \varepsilon_{i,k} \qquad (2.43)$$

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

which for I calibration samples ($\mathbf{R}_{I\times J}$) is:

$$\mathbf{c}_k = \mathbf{1}\beta_{0,k} + \mathbf{R}\beta_k + \varepsilon_k \qquad (2.44)$$

where $\mathbf{c}_k$ contains the concentration of the analyte k in these samples. By column-centering both $\mathbf{R}$ and $\mathbf{c}_k$ (substract the vector of the means of the columns of $\mathbf{R}$, $\bar{\mathbf{r}}$, from each sample response and substract the mean of the values of $\mathbf{c}_k$, $\bar{c}_k$, from all the calibration sample concentrations) the term $\beta_{0,k}$ is zero [3,45,63] and the model becomes (Figure 2.8):

$$\mathbf{c}_k = \mathbf{R}\beta_k + \varepsilon_k \qquad (2.45)$$

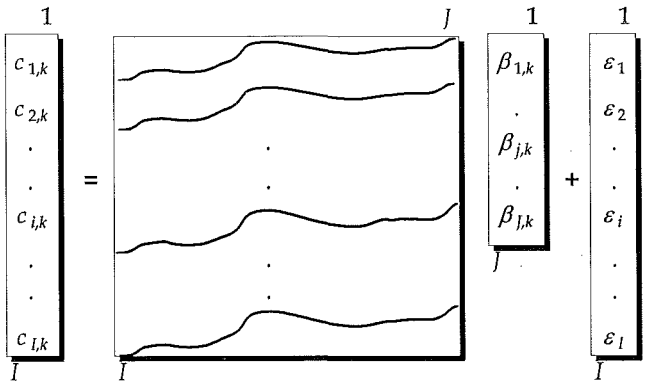


Figure 2.8. Matrix representation of eq 2.40

The estimation of β_k is:

$$\mathbf{b}_k = \mathbf{R}^+\mathbf{c}_k \qquad (2.46)$$

Each regression method calculates the pseudo-inverse matrix $\mathbf{R}^+$ in a different way which produces different estimated coefficients $\mathbf{b}_{k,\mathrm{ILS}}$, $\mathbf{b}_{k,\mathrm{PCR}}$ and $\mathbf{b}_{k,\mathrm{PLS}}$. To calculate $\mathbf{R}^+$, $\mathbf{R}$ is first decomposed into three matrices [3,43,47,49,64,65]:

$$\mathbf{R} = \mathbf{U}\mathbf{D}\mathbf{W}^T \qquad (2.47)$$

where the column vectors of $\mathbf{U}_{I\times Q}$ and $\mathbf{W}_{J\times Q}$ are orthonormal and $\mathbf{D}_{Q\times Q}$ can be either bidiagonal (PLS1) or diagonal (ILS,PCR) and Q is the rank of $\mathbf{R}$. For ILS and PCR, eq 2.47 is the singular-value decomposition (SVD) of $\mathbf{R}$ (see §2.4.2.3). PLS uses other algorithms for the decomposition [3,43,57,65]. The pseudo-inverse is calculated as [43,64]:

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

$$\mathbf{R}^{+} = \mathbf{W}\mathbf{D}^{-1}\mathbf{U}^{T} \qquad (2.48)$$

Before evaluating eq 2.48, PCR and PLS (not ILS) truncate the three matrices so that a low dimension approximation of the data is obtained which retains the relevant information and has less noise (see §2.4.2.2).

Prediction with inverse models

The predicted concentration of the analyte k in an unknown sample can be calculated either using raw data (eq 2.49), $\mathbf{r}_{un}$ centered (eq 2.50) or using both $\mathbf{r}_{un}$ and $c_{un,k}$ centered (eq 2.51):

$$c_{un,k} = b_{0,k} + b_{1,k} r_{un,1} + \ldots + b_{J,k} r_{un,J} = \bar{c}_k - \bar{\mathbf{r}}^{T}\mathbf{b}_k + \mathbf{r}_{un}^{T}\mathbf{b}_k \qquad (2.49)$$

$$c_{un,k} = \bar{c}_k + \mathbf{r}_{un,c}^{T}\mathbf{b}_k \qquad (2.50)$$

$$c_{un,k,c} = \mathbf{r}_{un,c}^{T}\mathbf{b}_k \qquad (2.51)$$

where $\mathbf{r}_{un,c} = \mathbf{r}_{un} - \bar{\mathbf{r}}$ and $c_{un,k,c} = c_{un,k} - \bar{c}_k$. The eq 2.50 is frequently found in the literature [42,43,62,66] and corresponds to models built with centered data and $\mathbf{r}_{un}$ centered before prediction. Eq 2.51 can be easily converted into eq 2.49 by uncentering the data: $c_{un,k} - \bar{c}_k = (\mathbf{r}_{un} - \bar{\mathbf{r}})^{T}\mathbf{b}_k$. If the constant term in eq 2.43 and 2.44 is not present [1,43,49,51,57,64,65,67,68], eqs 2.45 to 2.48 are used but without centering the data, and the prediction is given by eq 2.49 without the $b_{0,k}$.

The mathematical expressions for the variance of the estimated coefficients and of the predicted concentration in ILS, PCR and PLS can be found in Ref. 41,47, 58,62,.

2.4.2.1 Inverse Least Squares (ILS)

Inverse least squares (ILS), also known as P-matrix calibration, is a least-squares method that assumes the inverse calibration model given in eq 2.38. The error $\varepsilon_{i,k}$ is assumed to derive from uncertainties in the determination of the concentration in the calibration samples whereas no error is assumed in the absorbance values.

2.4.2.1.1 Calibration

Calibration consists of solving the system of I equations in eq 2.45 or eq 2.46. For both $\mathbf{R}$ and $\mathbf{c}_k$ column-centered, the coefficients are calculated as [43]

$$\mathbf{b}_{k,\mathrm{ILS}} = \mathbf{R}^{+}\mathbf{c}_k \tag{2.52}$$

$$b_{0,k} = \bar{c}_k - \bar{\mathbf{r}}^{\mathrm{T}}\mathbf{b}_k \tag{2.53}$$

The same $\mathbf{b}_{k,\mathrm{ILS}}$ is obtained in eq 2.52 if $\mathbf{c}_k$ contains the raw values[62]. $\mathbf{R}^{+}$ can be calculated as $(\mathbf{R}^{\mathrm{T}}\mathbf{R})^{-1}\mathbf{R}^{\mathrm{T}}$ or using the SVD of $\mathbf{R}$ (eqs 2.42 and 2.43). For $\mathbf{R}^{\mathrm{T}}\mathbf{R}$ to be invertible, it must be full rank. That is, all the columns and at least J rows (equal to the number of coefficients) must be linearly independent. Usually a higher number of calibration samples are analyzed to improve the precision of the estimated coefficients.

2.4.2.1.2 Prediction

Prediction from the response of the unknown sample $\mathbf{r}_{\mathrm{un}}$ is given by eqs 2.49, 2.50 or 2.51 depending on the preprocessing of the data.

Simultaneous ILS modeling of several components

The calibration and prediction can also be simultaneously done for K analytes as:

$$\mathbf{C} = \mathbf{1}\beta_0^T + \mathbf{R}\beta + \mathbf{E} \tag{2.54}$$

where $\mathbf{C}=[\mathbf{c}_1, \ldots, \mathbf{c}_K]$, $\beta=[\beta_1, \ldots, \beta_K]$, $\mathbf{E}=[\mathbf{e}_1, \ldots, \mathbf{e}_K]$ and $\beta_0=[\beta_{0,1}, \ldots, \beta_{0,K}]^T$. The least-squares solution for centered data is:

$$\mathbf{B} = \mathbf{R}^+\mathbf{C} \tag{2.55}$$

$$\mathbf{b}_0^T = \bar{\mathbf{c}}^T - \bar{\mathbf{r}}^T\mathbf{B} \tag{2.56}$$

where $\bar{\mathbf{c}}$ is the vector of the means of the columns of C. The predicted concentration of several analytes in the unknown sample is:

$$\mathbf{c}_{un}^T = \bar{\mathbf{c}}^T - \bar{\mathbf{r}}^T\mathbf{B} + \mathbf{r}_{un,c}^T\mathbf{B} \tag{2.57}$$

or, if the mean of the calibration set responses is substracted from $\mathbf{r}_{un}$ before prediction:

$$\mathbf{c}_{un}^T = \bar{\mathbf{c}}^T + \mathbf{r}_{un,c}^T\mathbf{B} \tag{2.58}$$

Prediction using the net analyte signal

Recently, Lorber *et al.*[69] described that the NAS of the calibration and unknown samples in ILS can be evaluated respectively as:

$$\mathbf{r}_{i,k}^* = (\mathbf{I} - \mathbf{R}_k^T(\mathbf{R}_k^T)^+)\mathbf{r}_i \tag{2.59}$$

$$\mathbf{r}_{un,k}^* = (\mathbf{I} - \mathbf{R}_k^T(\mathbf{R}_k^T)^+)\mathbf{r}_{un} \tag{2.60}$$

where $\mathbf{R}_k = \mathbf{R} - [\mathbf{c}_k\mathbf{r}^T/(\mathbf{r}^T\mathbf{R}^+\mathbf{c}_k)]$ and $\mathbf{r}$ in this expression is a linear combination of the rows of $\mathbf{R}$ which must include information about the spectrum of the analyte k. Then, the vector of sensitivities for the calibration sample i is: $\mathbf{s}_{i,k}^* = \mathbf{r}_{i,k}^*/c_{i,k}$, where $c_{i,k}$ is the

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

concentration in the *i*th calibration sample. In an errorless situation, all calibration samples should produce the same vector $\mathbf{s}_{i,k}^*$. This is not the case in real situations and the different $\mathbf{s}_{i,k}^*$ may be combined to form an estimate $\mathbf{s}_k^*$ (e.g. the mean value) that is representative of all the calibration set. Then, the concentration of the analyte *k* in the unknown sample is $c_{un,k}=\|\mathbf{r}_{un,k}^*\| / \|\mathbf{s}_k^*\|$.

2.4.2.1.3 Advantages of ILS

1. ILS may have some advantages over the methods based on latent variables such as PCR or PLS: ILS is easier to chemically interpret and it enables the computation of a higher number of statistical parameters, such as the statistical determination of confidence limits for the concentrations.

2.4.2.1.4 Limitations of ILS

1. The number of measured wavelengths is restricted to a subset of the available spectral wavelengths since the number of calibration samples must be equal or superior to the number of coefficients to be estimated. Using many wavelengths would require analyzing a large number of samples with the reference or well-established method which could be costly and tedious.

2. Collinear wavelengths must be avoided since they cause large variances for $\mathbf{b}_k$ and $c_{un,k}$ (see §2.6).

3. To build accurate ILS models, the constituent of interest must absorb at the selected wavelengths. The prediction improves if wavelengths related to constituent of interest are added to the model. However, too many wavelengths may include spectral noise which is unique to the training set and degrade the prediction accuracy for unknown samples that are more unlikely to vary in exactly the same manner (the *overfitting* problem). Ideally, there is a crossover point between selecting enough wavelengths to compute an accurate least squares solution and selecting few enough so that the calibration is not affected by the collinearity of the spectral data.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

4. ILS has a more reduced signal averaging effect than CLS since just a few responses are considered[43].

Selecting the appropriate set of wavelengths is critically important for the final quality of the ILS model. This can be done using the chemical and spectral knowledge about the analyte of interest and the interferents. In absence of this information, empirical selection methods such as stepwise regression [70] or genetic algorithms[71,72] based on some quality criterion can be used for this purpose. Other alternatives are the regression techniques which can handle collinear data, such as ridge regression (RR) (although RR may solve the collinearity problem but does not reduce the number of wavelengths used) or factor-based methods such PCR and PLS that use linear combinations of all the wavelengths. A recent study compared the PCR, RR and PLS[73].

2.4.2.2 Factor-based regression methods (PCR and PLS)

Principal component regression (PCR) and partial least squares (PLS) are factor-based regression methods that solve some limitations of CLS and ILS. Both PCR and PLS express the correlated information in the many measured variables in a new coordinate system of a few "latent" variables (*factors*) that are a linear combination of the original variables. This is done by decomposing the matrix of instrumental responses of the I calibration samples $\mathbf{R}_{I\times J}$ (often column-centered or autoscaled) into the product of two smaller matrices[51]:

$$\mathbf{R}_{I\times J} = \mathbf{t}_1 \mathbf{p}_1^T + \mathbf{t}_2 \mathbf{p}_2^T + \ldots + \mathbf{t}_A \mathbf{p}_A^T + \mathbf{E} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \qquad (2.54)$$

where $\mathbf{T}_{I\times A}=[\mathbf{t}_1,\ldots, \mathbf{t}_A]$ (*scores*) and $\mathbf{P}_{J\times A}=[\mathbf{p}_1,\ldots, \mathbf{p}_A]$ (*loadings*) are full rank matrices, A is the number of factors and $\mathbf{E}_{I\times J}$ is the part of the data that is not modeled (Figure 2.9). The ith row of $\mathbf{T}_{I\times A}$, $\mathbf{t}^i$, contains the coordinates of the ith calibration sample in this new system, called *scores*. The ath column $\mathbf{t}_a$ contains the scores of all the samples in the factor a.. Each score is a linear combination of the instrumental measurements:

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

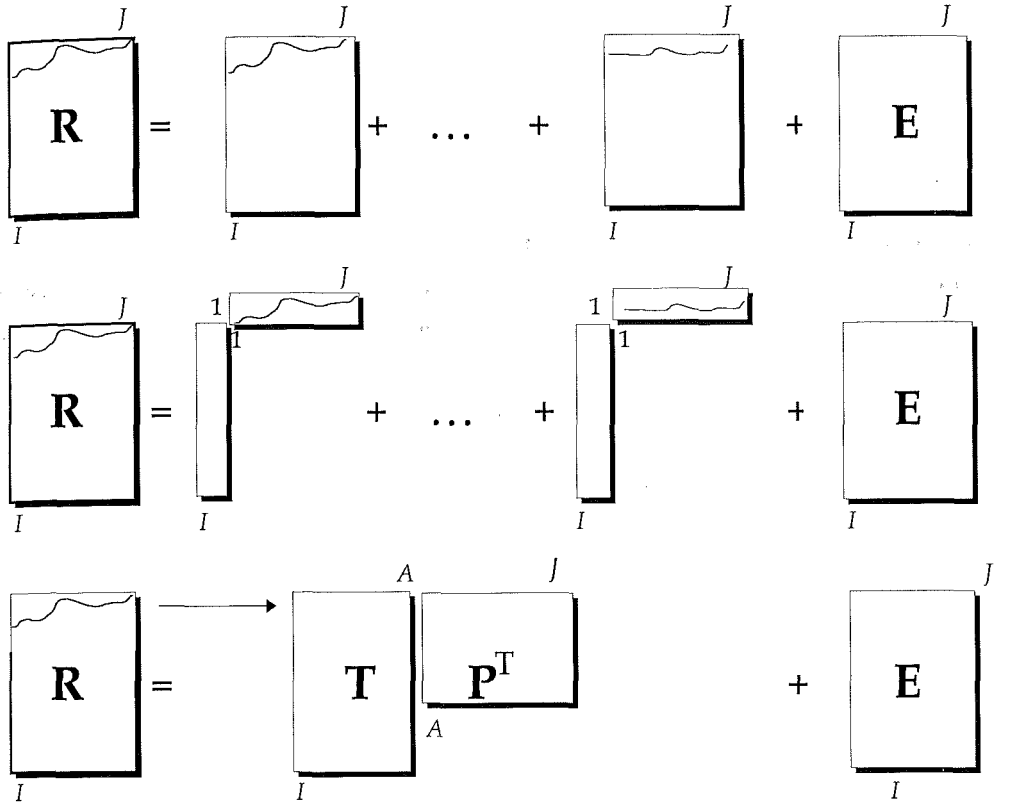


Figure 2.9. Example of the factor decomposition of a matrix of instrumental responses (here spectra).

$$t_{i,a} = w_{a,1}\, r_{i,1} + w_{a,2}\, r_{i,2} + \ldots + w_{a,J}\, r_{i,J} \qquad (2.62)$$

where the weighting coefficients $w_{a,j}$ are found during calibration. In PCR, $w_{a,j} = p_{a,j}$, the elements of **P**. The scores are unique to each calibration spectrum and are used instead of the original responses as variables that represent the sample in the regression with the concentration. In general, the assumed model for these methods is of the form[74]:

$$c_{i,k} = \theta_{0,k} + \theta_{1,k}\, t_{i,1} + \ldots + \theta_{a,k}\, t_{i,a} + \ldots + \theta_{A,k} t_{iA} + f_{i,k} = \theta_{0,k} + \mathbf{t}^{i\,T}\theta_k + f_{i,k} \qquad (2.63)$$

where $t_{i,a}$ is the ath score of the ith sample and $\theta_{k,a}$ are the model coefficients.

A (called the *pseudo-rank* of the calibration matrix) is the number of factors that are important for regression (and is usually much smaller than the number of measurements). It is supposed that $\mathbf{T}_{I\times A}\mathbf{P}_{J\times A}^{T}$ describes the part of **R** that comes from changes in the chemistry and is related to the constituents of interest, while $\mathbf{E}_{I\times J}$ contains information non important for predicting concentration as, for example, random noise. By retaining *A* factors, $\mathbf{R}_A=\mathbf{T}_{I\times A}\mathbf{P}_{J\times A}^{T}$ is an approximate reproduction of **R** with a lower mathematical rank that retains the non-random sources of variation and contains less noise. Thus the data compression step in these techniques is a way of filtering out noise, which is distributed throughout all loading vectors while the true spectral variation is generally concentrated in the early loading vectors. Discarding a part of the data introduces bias in the estimated coefficients but decreases their variance, thus improving the predictive ability of the model. For this reason PCR and PLS are called biased methods[43].

An infinite set of decompositions obeys eq 2.61 and different constrains distinguish the different regression methods that give a different new coordinate system[41]. In PCR (and usually in PLS) the score vectors $\mathbf{t}_a$ are orthogonal to each other. The additional constraint of orthonormal columns for **P** gives the principal component decomposition while **P** is not orthogonal in the other decompositions[63].

T, **P**, the model coefficients and the model size *A* are found during the calibration. If the score vectors are orthogonal, the resulting parameter estimates $q_{k,a}$ are stable. The prediction for a new sample is given by:

$$c_{un,k} = q_{0,k} + q_{1,k}\,t_{un,1} + \ \ldots\ + q_{a,k}\,t_{un,a} + \ldots + q_{A,k}\,t_{unA} = q_{0,k} + \mathbf{t}_{un}^{T}\mathbf{q}_k \qquad (2.64)$$

where $q_{k,a}$ are the estimations of $\theta_{k,a}$ and $t_{un,a}$ are the scores of the response of the unknown sample $\mathbf{r}_{un}$ calculated with eq 2.56. The prediction equation, expressed as a function of the scores, can be converted into eq 2.20, as a function of the measured responses. The PCR and PLS coefficients can also be found with eq 2.41, where $\mathbf{R}^{+}$ is calculated from $\mathbf{R}_A$, the matrix reproduced with only the significant factors which are supposed to be related to the chemical signal[52] and the factors associated mostly with random errors are not used.

Estimation of the optimal number of factors

Unlike CLS or ILS that calculate only one model, PCR and PLS models can be built with a different number of factors. However, an unnecessary large number leads to *overfitting* and bad prediction due to the inclusion of factors that model noise; and a too small number (*underfitting*) introduces systematic error since not enough terms are used to model all the spectral variations of the constituents of interest. Among others [75], the usual method of selecting the optimal number of factors A is to build models with an increasing number of factors and measure their predictive ability using samples of known concentration [43-45,47,73,76]. Then A corresponds to the first local minimum of the plot number of factors versus predictive ability. Statistical tests for determining if including an additional factor is significant have been also described [77,51]. The predictive ability is usually evaluated as $\text{PRESS}=\sum_i (c_i - c_{i,pred,a})^2$ (Prediction Residual Error Sum of Squares) where c_i and $c_{i,pred,a}$ are, respectively, the measured and predicted analyte concentration with a factors in the sample i). The way $c_{i,pred,a}$ is found defines two different validation methods[47, 78]:

- *External validation*. The $c_{i,pred,a}$ values are predicted from a set of I_P samples (*validation set*) not used in the model-building step and measured under the same conditions as the training set. The model with a factors used for prediction has been built using the training set (Figure 2.10) The predictive ability of the model is given by the *Root-Mean-Square Error of Prediction* (RMSEP) defined as:

$$\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^{I_P} (c_i - c_{i,pred,a})^2}{I_P}} \qquad (2.65)$$

where c_i and $c_{i,pred,a}$ are, respectively, the measured and predicted analyte concentration with a factors in the sample i . RMSEP is equivalent to PRESS but it may be preferable since it has the same units as the concentration values. As I_P gets large, the RMSEP will approach the prediction error of the population of all future samples. The test set validation gives the best estimate of a model's performance since none of the samples in the validation set is used in the model building and

the final calibration equation is used for prediction. Its drawback is that, to reliably estimate the prediction ability, the data set must be representative for the future unknown samples and cover the expected range of concentration values. This may require a large number of test objects. Moreover it is rather wasteful and expensive due to the time and cost involved in generating the independent data set which is used for testing purposes only. Some alternative techniques overcome these problems and use all the available data both for calibration and for testing. Only cross-validation is considered here.

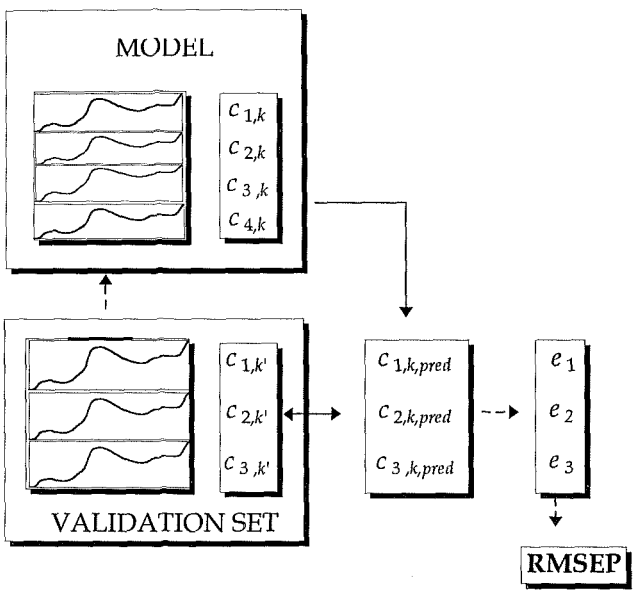


Figure 2.10 Graphical representation of test set validation.

In the *cross-validation* method, part of the data is left out, a model is constructed using the remaining data, and a prediction is made on the left-out data. This process is repeated until all the samples have been left out once. The most used is *leave-one-out cross-validation* where each sample is left out one at a time. An approximation to the prediction error is given by the *Root-Mean-Square Error of Cross-validation* (RMSECV):

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

$$\text{RMSECV} = \sqrt{\frac{\sum_{i=1}^{I}(c_i - c_{i,pred,a})^2}{I}} \tag{2.66}$$

where I is the number of calibration samples and $c_{i,pred,a}$ is the predicted analyte concentration when a model with a factors is constructed without sample i. Since the predicted samples are not used to build the model, RMSECV is a good approximation to the prediction error for unknown samples and the model is validated without having to measure an entirely new set of data. Cross-validation is used also when it is difficult to obtain enough prediction samples to make RMSEP significant. In addition, since each sample is left out of the model during cross-validation, outliers can be detected by comparing the spectral reconstruction to the original training spectrum or the concentrations. A drawback of leave-one-out cross-validation is the large time required since the model is re-calculated for every sample left out. In such cases, leaving out groups of samples at a time can be preferable. This is also used in training sets that contain replicate spectra of the same sample.

It must be also said that the most desirable situation in multivariate calibration would be to have enough data to divide it into three sets: calibration set (to calculated the model), validation (or monitoring) set, to determine the optimal number of factors, and test (or prediction) set, an independent set not used to determine the model and that evaluates the final quality of the model[74]. However, when the number of samples is not so large to enable this division, the model is constructed with the calibration set, the factors selected with cross-validation (thus using the calibration set) and finally validated with the test set.

2.4.2.2.1 Advantages of factor-based regression methods (PCR and PLS)

Factor-based models combine the best features of both the CLS and ILS methods and are generally better in both accuracy and robustness:

1. Both PCR and PLS decompose the original matrix into factors and retain a subset of them to calculate the regression equation. This produces robust models for predicting concentrations of the desired constituents in very complex samples even that may contain contaminants not present in the original calibration mixtures.

2. The number of calibration samples, that in ILS must be at least equal to the number of wavelengths, in PCR and PLS is not necessarily large for mathematical requirements since only the coefficients for a reduced number of factors must be calculated.

3. Unlike in ILS, wavelength selection is not necessary in PCR and PLS. Usually the whole spectrum, or wide regions are used. This gives the signal averaging effect of full-spectral technique such as CLS and, together with the factors decomposition, makes models less susceptible to spectral noise.

4. The collinearity problem met in ILS is eliminated since the scores are usually orthogonal.

2.4.2.2.2 Limitations of factor-based regression methods (PCR and PLS)

1. Although PCR and PLS are full-spectrum methods and are able to accommodate some degree of non-linearities, the inclusion of non-informative wavelengths or that have non-linearities can degrade performance[44]. A careful wavelength selection is advisable to improve the prediction ability of these models.

2. Models are more complex to understand and interpret than the CLS and ILS methods and calculations are slower.

3. The method for selecting the important factors is a critical step in these methods and has been the subject of a large number of papers.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

2.4.2.3 Principal component analysis (PCA)

Principal component analysis (PCA) [41,79]constitutes one of the most important methods used in the chemometric literature and the factor-decomposition technique used in PCR. PCA transforms the correlated variables in the calibration data to a new set of A uncorrelated variables according eq 2.61 (without loss of generality $\mathbf{R}_{I\times J}$ is assumed to be mean-centered [62,63,79]:

$$\mathbf{R}_{I\times J} = \mathbf{t}_1\mathbf{p}_1^T + \ldots + \mathbf{t}_a\mathbf{p}_a^T + \ldots + \mathbf{t}_A\mathbf{p}_A^T + \mathbf{E} = \mathbf{T}_A\mathbf{P}_A^T + \mathbf{E} \qquad (2.61)$$

where the matrix of *scores* $\mathbf{T}_A=[\mathbf{t}_1, \ldots, \mathbf{t}_a, \ldots, \mathbf{t}_A]$ $(I\times A)$ has orthogonal columns ($\mathbf{t}_a^T\mathbf{t}_b=0$, $a\neq b$), the matrix of *loadings* $\mathbf{P}_A=[\mathbf{p}_1, \ldots, \mathbf{p}_a, \ldots, \mathbf{p}_A]$ $(J\times A)$ is orthonormal ($\mathbf{p}_a^T\mathbf{p}_b=0$, $a\neq b$ and $\|\mathbf{p}_a\|^2=\mathbf{p}_a^T\mathbf{p}_a=1$) and $\mathbf{E}_{I\times J}$ is the part of **R** not retained by the decomposition. The columns in **T** are ordered in decreasing order of explained variance so that the first factors (and thus a reduced number of variables) express the most important information in the data without a significant loss. By introducing normalized scores $\mathbf{u}_a=\mathbf{t}_a/\lambda_a^{1/2}$, eq 2.54 can be written as [63,80]:

$$\mathbf{R}_{I\times J}=\lambda_1^{1/2}\mathbf{u}_1\mathbf{p}_1^T + \ldots + \lambda_a^{1/2}\mathbf{u}_2\mathbf{p}_2^T + \ldots + \lambda_A^{1/2}\mathbf{u}_A\mathbf{p}_A^T + \mathbf{E} = \mathbf{U}_A\mathbf{D}_A\mathbf{P}_A^T + \mathbf{E} \qquad (2.67)$$

where $\lambda_a=\mathbf{t}_a^T\mathbf{t}_a$ is the square norm of the ath score vector prior to normalization and it is also the ath eigenvalue of $\mathbf{R}^T\mathbf{R}$, $\mathbf{D}_A=\text{diag}(\lambda_a^{1/2})$ $(A\times A)$ is diagonal and $\mathbf{U}_A=[\mathbf{u}_1,\ldots,\mathbf{u}_a,\ldots,\mathbf{u}_A]=\mathbf{T}_A\mathbf{D}_A^{-1}$ $(I\times A)$ is orthonormal and contains the vectors $\mathbf{t}_a$ normalized to length one. $\mathbf{R}_A=\mathbf{U}_A\mathbf{D}_A\mathbf{P}_A^T=\mathbf{T}_A\mathbf{P}_A^T$ is the compressed representation of **R** with some of the noise removed. The NIPALS algorithm[41,51] decomposes **R** according to eq 2.61. The singular-value decomposition (SVD)[1, 43,54] of **R** gives $\mathbf{R}=\mathbf{UDP}^T$ so that $\sigma_a=\lambda_a^{1/2}$ are the singular values of **R** and $\sigma_1\geq\sigma_2\geq\ldots\geq\sigma_A\geq 0$.The matrices $\mathbf{U}_A$, $\mathbf{D}_A=\text{diag}(\sigma_1,\ldots\ \sigma_A)$ and $\mathbf{P}_A$ are the retained part from partitioning the matrices $\mathbf{U}=[\mathbf{U}_A, \mathbf{U}_{-A}]$, $\mathbf{P}=[\mathbf{P}_A,\mathbf{P}_{-A}]$ and $\mathbf{D}=\begin{bmatrix}\mathbf{D}_A & \mathbf{0}\\ \mathbf{0} & \mathbf{D}_{-A}\end{bmatrix}$ (**0** is a matrix of zeros). $\mathbf{P}_A$ is the updated **P** matrix without the columns corresponding to irrelevant factors. It can also be seen that $\mathbf{D}_A^2=\mathbf{T}_A^T\mathbf{T}_A$

2.4.2.3.1Loadings

The loadings are the elements of the arbitrarily[79] normalized eigenvectors of $\mathbf{R}^T\mathbf{R}$, $\mathbf{p}_a$, with associated eigenvalue λ_a. Since $\mathbf{R}^T\mathbf{R}$ is a symmetric matrix, its eigenvectors are orthogonal. The equation of eigenvalues of $\mathbf{R}^T\mathbf{R}$ is:

$$\mathbf{R}^T\mathbf{R}\mathbf{p}_a=\mathbf{p}_a\lambda_a \tag{2.68}$$

and for the complete set of eigenvectors $\mathbf{P}=[\mathbf{p}_1,...,\mathbf{p}_A]$:

$$\mathbf{R}^T\mathbf{R}\mathbf{P} = \mathbf{P}\text{diag}(\lambda_a) \tag{2.69}$$

where $\text{diag}(\lambda_a)$ is the diagonal matrix of eigenvalues. By multiplying both sides of eq 2.69 by $\mathbf{P}^T$, the resulting equation

$$\mathbf{R}^T\mathbf{R} = \mathbf{P}\text{diag}(\lambda_a)\mathbf{P}^T \tag{2.70}$$

shows that $\mathbf{P}$ diagonalizes $\mathbf{R}^T\mathbf{R}$ associated with the eigenvalues $\text{diag}(\lambda_a)$, which agrees with $\mathbf{R}^T\mathbf{R}=(\mathbf{U}\mathbf{D}\mathbf{P}^T)^T(\mathbf{U}\mathbf{D}\mathbf{P}^T)=\mathbf{P}\mathbf{D}^T\mathbf{U}^T\mathbf{U}\mathbf{D}\mathbf{P}^T=\mathbf{P}\mathbf{D}^2\mathbf{P}^T$ where $\mathbf{D}^2=\text{diag}(\lambda_a)$.

2.4.2.3.2 Scores

Each vector of scores $\mathbf{t}_a$ is the projection of $\mathbf{R}$ on the basis vector $\mathbf{p}_a$ calculated as $\mathbf{t}_a = \mathbf{R}\mathbf{p}_a$. The scores on the first A principal components are $\mathbf{T}_A=\mathbf{R}\mathbf{P}_A=\mathbf{U}_A\mathbf{D}_A$. It is not necessary to use the inverse (or pseudo inverse) of $\mathbf{P}$ to solve the equation $\mathbf{R}=\mathbf{T}\mathbf{P}^T$ since the matrix $\mathbf{P}$ is orthonormal and it is enough to multiply both sides of this equation for $\mathbf{P}$. If $\mathbf{R}$ has been centered (thus $\mathbf{R}^T\mathbf{R}/(I\text{-}1)$ is a covariance matrix), the resulting scores are centered; otherwise, the scores are not centered. The scores for a new object of coordinates $\mathbf{r}_i^T$ are $\mathbf{t}^i{}_A{}^T=\mathbf{r}_i^T\mathbf{P}_A$.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

2.4.2.3.3 Eigenvalues

The sum of the eigenvalues is equal to the total variance of the data set and to the trace of the original matrix. Each eigenvalue λ_a divided by the trace represents the proportion of the total variance accounted for by the eigenvector $\mathbf{p}_a$.

2.4.2.3.4 Number of significant factors

Criteria such as the empirical indicator function by Malinowski[81] or a PRESS value using cross-validation [43,79,82], among others[83] have been proposed for selecting the number of factors that explain a significant variance of **R**.

2.4.2.3.5 Advantages of PCA

1. A large number of original variables can be reduced to few new variables that account for a significant portion of the information (variance) of the data. The reduced data can be interpreted as primary sources of variation of the original data. The eigenvectors which model statistically significant variation in the data are retained. This allows the graphical representation of the samples in the reduced space with a minimum loss of information, to identify natural associations of samples and/or variables and their relationship as well as outlier detection[79].

2. By deleting the principal components whose eigenvalues are nearly zero, linear dependencies are removed. If these PCs are associated to noise in the data, the noise in the reproduced data matrix has been reduced.

2.4.2.4 Principal component regression (PCR)

2.4.2.4.1 Calibration

PCR creates a quantitative model in a two-step process: the PCA scores $\mathbf{T}_A$ of I calibration samples are calculated for A factors and then the scores are regressed against the analyte concentration:

$$\mathbf{c}_k = \mathbf{1}\,\theta_{0,k} + \mathbf{T}_A\,\theta_k + \varepsilon \tag{2.71}$$

where θ_k is the vector of the regression coefficients and ε is a vector of independent and normally distributed errors. For column-centered data (i.e. the average calibration spectrum is substracted from each spectrum, and the average calibration concentration is substracted from each concentration) the resulting scores are centered and the intercept is eliminated from the fit (Figure 2.11):

$$\mathbf{c}_k = \mathbf{T}_A\theta_k + \varepsilon_k \tag{2.72}$$

$$\underset{I}{\begin{bmatrix} c_{1,k} \\ c_{2,k} \\ \cdot \\ c_{i,k} \\ \cdot \\ c_{I,k} \end{bmatrix}} = \underset{I}{\overset{A}{\begin{bmatrix} t_{1,1} & \cdot & t_{1,A} \\ t_{2,1} & \cdot & t_{2,A} \\ \cdot & \cdot & \cdot \\ t_{i,1} & \cdot & t_{i,A} \\ \cdot & \cdot & \cdot \\ t_{I,1} & \cdot & t_{I,A} \end{bmatrix}}} \underset{A}{\overset{1}{\begin{bmatrix} \theta_{1,k} \\ \cdot \\ \theta_{A,k} \end{bmatrix}}} + \underset{I}{\overset{1}{\begin{bmatrix} \varepsilon_{1,k} \\ \varepsilon_{2,k} \\ \cdot \\ \varepsilon_{i,k} \\ \cdot \\ \varepsilon_{I,k} \end{bmatrix}}}$$

Figure 2.11 Matrix representation of eq 2.72

Eq 2.72 can also be deduced by considering a model for centered data[73]:

$$\mathbf{c}_k = \mathbf{R}\,\beta_k + \varepsilon \tag{2.73}$$

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

and multiplying by $\mathbf{P}_A\,\mathbf{P}_A^T=\mathbf{I}$:

$$\mathbf{c}_k = \mathbf{R}\,\mathbf{P}_A\,\mathbf{P}_A^T\,\beta_k + \varepsilon \qquad (2.74)$$

which gives eq 2.72 where $\mathbf{T}_A= \mathbf{R}\,\mathbf{P}_A$, $\theta_k= \mathbf{P}_A^T\,\beta_k$ and $\mathbf{P}_A$ only contains the loadings for the A significant factors. The least-squares solution for θ_k has the same form as the ILS solution but with $\mathbf{T}$ and $\mathbf{q}_k$ instead of $\mathbf{R}$ and $\mathbf{b}_k$:

$$\mathbf{q}_k = \mathbf{T}_A{}^{+}\mathbf{c}_k \qquad (2.75)$$

The coefficients $\mathbf{b}_k$ in eq 2.20 can be calculated from $\mathbf{q}_k$ as:

$$\mathbf{b}_{k,PCR}=\mathbf{P}_A\,\mathbf{q}_k \qquad (2.76)$$

$\mathbf{b}_{k,PCR}$ can also be estimated [47] from eqs 2.46 to 2.48 with $\mathbf{R}_A{}^{+}=\mathbf{P}_A\mathbf{D}_A{}^{-1}\mathbf{U}_A{}^T$ (proof: $\mathbf{b}_k= \mathbf{P}_A\mathbf{q}_k= \mathbf{P}_A\mathbf{T}^{+}\mathbf{c}_k= \mathbf{P}_A(\mathbf{T}_A{}^T\mathbf{T}_A)^{-1}\mathbf{T}_A{}^T\mathbf{c}_k =\mathbf{P}_A\mathbf{D}_A{}^{-2}\mathbf{D}_A\mathbf{U}_A\mathbf{c}_k =\mathbf{P}_A\mathbf{D}_A{}^{-1}\mathbf{U}_A{}^T\mathbf{c}_k =\mathbf{R}_A{}^{+}\mathbf{c}_k$ since $\mathbf{T}_A{}^T\mathbf{T}_A=\mathbf{D}_A{}^2$; $\mathbf{T}_A=\mathbf{U}_A\mathbf{D}_A$; $\mathbf{R}_A=\mathbf{U}_A\mathbf{D}_A\mathbf{P}_A{}^T$)

2.4.2.4.2 Prediction

The concentration of the analyte k in an unknown sample whose response is $\mathbf{r}_{un}$ can be predicted in two equivalent ways:

a) simplified prediction, using equations 2.49, 2.50 or 2.51 with the coefficients given by eq 2.76.

b) full prediction, using the loading vectors $\mathbf{P}_{J\times A}$ to transform $\mathbf{r}_{un}$ (centered) to its factor scores $\mathbf{t}_{un,A}{}^T= \mathbf{r}_{un}{}^T\mathbf{P}_A$ (Figure 2.12):

$$c_{un,k} = \bar{c} + \mathbf{t}_{un,A}{}^T\,\mathbf{q}_k \qquad (2.77)$$

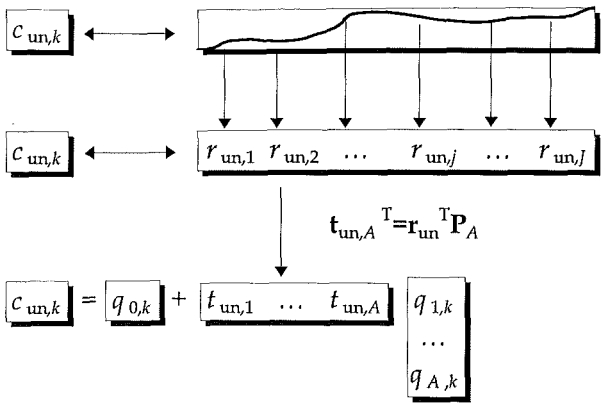


Figure 2.12. Matrix representation of the full prediction in PCR (eq 2.77).

2.4.2.4.3 Selection of factors in PCR

The usual procedure for selecting the optimal number of factors A in PCR is to calculate the predictive ability of several models built with a different number of factors (see §2.4.2.2). The factors can be included in the models in two different ways: (a) in decreasing order of explained variance of the response data matrix (i.e. factors ordered by magnitude of their eigenvalues, the *top-down* approach) or (b) ordered according to the importance for predicting each analyte. The most used strategy is (a). The PCs are calculated independently of any knowledge of the analyte concentrations and merely represent the largest common variations in the response data (e.g. spectra). Presumably, the major variations within the independent variables, accounted for in the PCs with the largest eigenvalues, are related to changes in the constituent concentrations. So, only these PCs are retained. The noise, which usually provides the smallest contribution to the data, is supposed associated to the factors with the smallest variances (that is, with the smallest eigenvalues). These factors are probably irrelevant for the prediction of chemical compositions and are rarely used in regression[84,85].

The approach (b) is justified by the fact that the underlying effects that are directly related to the concentration of the constituent of interest may be small in comparison with the irrelevant ones and may not appear among the PCs which explain a large percentage of the variance of the data. Then the PCR model built

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

with the strategy (a) includes unnecessary factors not related to the constituent of interest that may degrade the prediction ability[86] of the model. It has been shown that PCR with selection of principal components instead of the usual top-down approach yields simpler and better models [86,87]. One way of determining the factors that are relevant for prediction is to build models with the factors included in all possible orders. However, this combinatorial problem can take a long computation time when the number of factors to be considered is large. The generalized simulated annealing (GSA) algorithm with the PRESS criterion [87] and a forward selection procedure where the PCs with the largest absolute correlation with the dependent variable enter one at a time in the model[86] have been used. Jouan-Rimbaud *et al.*[86] suggested that since the PCs are not correlated, simpler selection methods could probably be applied, but this required further investigations that were not made. Sun[84] presented the correlation principal component regression although no indications were given on how to perform the selection of the more relevant factors. Recently, Xie and Kalivas[88,89] proposed a forward selection procedure for PCR.

2.4.2.4.4 Advantages of PCR

Besides the advantages indicated in §2.4.2.2.1, the following can be noted:

1. The PCA data compression extracts the underlying effects in the **R** data and PCR uses these in the inverse regression to calculate the model coefficients and to predict the values of the dependent variable.

2. PCR combines the advantage of using all the spectral channels, excludes noise effects (which are relegated to the unused factors), and retains the ILS independence of uncalibrated components.

3. The problems present in collinear data such as NIR spectra have disappeared because the columns of **T** are orthogonal and the PCs of the smallest eigenvalues (the ones which would produce the largest variance in the estimated coefficients) have been deleted. This produces more reliable estimates of the model coefficients and hence a good predictive model useful for calibration of spectroscopic instruments.

2.4.2.4.5 Limitations of PCR.

The PCR calibration model is not completely free of problems:

1. The PCA factors are calculated independently of any knowledge of the concentration of the analyte of interest. The usual top-down method can introduce irrelevant factors that degrade the predictive ability of the model.

2.4.2.5 Partial least squares regression (PLS)

Partial least squares regression (PLS) is a factor-based calibration technique that uses both spectral and constituent concentration information in the decomposition process to find those factors with the greatest relevance for prediction. This is different from PCR, that first decomposes the spectral matrix into factors that represent the most common variations in the response data, completely ignoring their relation to the constituents of interest and then regresses the scores against the concentrations. The resulting PLS factors are more relevant for description of the concentration information than those calculated in PCR.

2.4.2.5.1 Calibration

The calibration equations for PLS are more complex than those of PCR and are not described here. Different versions of PLS algorithms can be found in a large number of publications [3,41,51,57,65,90-97]. Two main types of PLS algorithms exist: PLS-1, that calibrates for one constituent at a time, and PLS-2, that calibrates for more than one constituent simultaneously. The comparison of both methods can be found in Ref. 41.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

2.4.2.5.2 Prediction

The predicted concentration of the analyte k in an unknown sample whose response is r_{un} can be found using the eq 2.49 to 2.51, where the coefficients have been calculated using the PLS algorithm (see also Ref. 41 for more information of the prediction equations).

2.4.2.5.3 Advantages of PLS

1. Single step decomposition and regression; factors are directly related to constituents of interest rather than largest common spectral variations.

2. Calibrations are generally more robust provided that calibration set accurately reflects range of variability expected in unknown samples.

3. Enjoys the signal average advantages of other full-spectrum methods such as PCR and CLS [51].

2.4.2.5.4 Limitations of PLS

1. Models are more difficult to understand and interpret that CLS, ILS or PCR.

2. Calculations in PLS-1 are slower than most classical methods.

In the previous sections, four multivariate regression models have been presented. The selection of the samples and sensors used to build and validate the model is an important step that influences the quality of the predictions. In the following section, some ideas are given about the methodological selection of the best samples and sensors for calibration.

2.5 Optimal design in multivariate calibration

Calibration relates analyte concentrations and instrumental responses with the aim of achieving an acceptable predictive quality (in terms of trueness and precision) over all the experimental domain and applicable to the largest number of unknown samples possible. The mathematical expressions of the multiple linear regression (MLR) model, CLS, ILS and PCR are compared in the Table 2.1 along with their least-squares solution.

Table 2.1. Comparison of different regression models

Model	*Model expression*	*Least-squares solution*	*Vector of dependent variables*	*Matrix of the model*		
					Rows are:	*Columns are:*
MLR	$\mathbf{y} = \mathbf{X}\beta + \varepsilon$	$\mathbf{b} = \mathbf{X}^{+}\mathbf{y}$	$\mathbf{y}$	$\mathbf{X}$	experiments	variables
CLS	$\mathbf{r}_{un} = \mathbf{S}\,\mathbf{c}_{un,true} + \varepsilon$	$\mathbf{c}_{un} = \mathbf{S}^{+}\,\mathbf{r}_{un}$	$\mathbf{r}_{un}$	$\mathbf{S}$	sensors	pure spectra
ILS	$\mathbf{c}_k = \mathbf{R}\beta_k + \varepsilon$	$\mathbf{b}_k = \mathbf{R}^{+}\mathbf{c}_k$	$\mathbf{c}_k$	$\mathbf{R}$	samples	sensors
PCR	$\mathbf{c}_k = \mathbf{T}\theta_k + \varepsilon$	$\mathbf{q}_k = \mathbf{T}^{+}\mathbf{c}_k$	$\mathbf{c}_k$	$\mathbf{T}$	samples	scores

It can be seen that the dependent variables **y** in MLR are either the spectrum of the unknown sample $\mathbf{r}_{un}$ (CLS) or the concentration of the analyte under study in the calibration samples $\mathbf{c}_k$ (ILS, PCR). A row of the calibration matrix **X** corresponds either to the absorbances of the *K* analytes at a given wavelength (CLS) or to one calibration sample represented by the absorbances at *J* wavelengths (ILS) or its scores (PCR). A column of the calibration matrix (the settings of one variable in all the experimental points) can be either the spectrum of one pure analyte (CLS), the absorbance at a given wavelength in all the calibration samples (ILS) or the score of these samples in a given factor (PCR). The coefficients of the MLR model correspond to the concentration of the *K* analytes in the unknown sample in CLS.

The variance of the predicted concentration in these models depends on:

a) three sources of error: the measured responses from the unknown sample, measured responses from the calibration samples and the analyte concentrations in the calibration samples.
b) the mathematical expression of the model.
c) the points in the calibration matrices (the calibration space)
d) the position in the calibration space of the sample to be predicted (if the point is close or away from the points used for calibration).

The degree of complexity of the available expressions to calculate this variance in the different calibration models[1,41,47,58,62,98] depends on the assumptions made referent to the points a) to d). An usual simplification is to assume that the errors in the independent variables are neglected and to use the MLR expressions to calculated the variances. In this case, $var(c_{un,k})$ in ILS is given by eq 2.17 and only depends on the values in **R** (the absorbances of each calibration sample) and not on the values of the concentrations. In the same way, $var(\mathbf{c}_{un})$ in CLS depends only on the variance of the measured responses in $\mathbf{r}_{un}$ and on the matrix **S** (eq 2.12). With these assumptions, the DOE can be applied for the optimal building of the multivariate calibration models so that a reliable estimation of the searched relationship is at the minimum cost found[99].

As already indicated in §2.3.4, the values in the calibration matrices in MLR influence in the way that the measurements errors propagate to the predicted concentration. The DOE [16,21] indicates the most appropriate settings for the variables in each row of **X** to find a reliable estimation of the coefficients. However, classical designs (e.g. factorial-type designs) require the independent variables be manipulated according to the specified design strategy. This, for example, would require preparing calibration samples with specific values of absorbance at each of the measured wavelengths in ILS or scores in PCR and in CLS having pure analytes with the necessary absorbance values at each wavelength. This is not so readily used in multivariate calibration problems where it is impossible to make calibration samples of a determined composition and with prearranged values of the independent variables (specially when they are spectral measurements which are function of the chemical values). These situations are frequent in the quantitative analysis of natural products (i.e. water, flour, meat,...), whose chemical composition

cannot be controlled. Then the design variables in **S**, **R** or **T** are interrelated and cannot be manipulated independently of one another since they are influenced by many factors outside the experimenter's control, relating to the nature of the substance. These cases also arise in structure-activity relationship studies, where the values of the independent variables are fixed for the chemical configuration.

Therefore, the classical designs are difficult to use due to the impossibility of preparing samples with complex matrices and well determined instrumental responses. The alternative consists of obtaining a large list of possible points (rows in **S**, **R** or **T**) and select among them the most appropriate for calibration (generally the more economical subset that has the sufficient information for the model). This means that the best wavelengths(rows of **S**) are selected in CLS from the full spectra of the pure components (the matrix **S**). This is an advantage in ILS and PCR since the best calibration samples can be selected from a series of all the available candidate samples characterized by "inexpensive" multivariate measurements (e.g. spectroscopic data which can be collected with little labor and cost, **R**) or scores (**T**). For calibration, the analyte concentration must only be determined in the few selected samples using the more-time consuming referee or well-established method. Compared to analyzing a full set of samples, the cost of the model is being reduced.

The requirements for this selection step have been indicated in §2.3.4. The selection algorithms select the subset of points among all the points available for multiple linear regression (MLR) optimizing the desired criterion. An adequate criterion is the optimality of the selected subset for estimating the parameters of the model, since a low variance in the parameters should result in a good predictive power over the calibration range (e.g. the D-criterion and the A-criterion). Other criterion to consider is the quality of the predictions furnished by the model built (the G-criterion). Also measures of collinearity must be considered, specially for the selection of samples in ILS. The collinearity in the columns of **R** produces large uncertainty in the coefficients in the model and thus a large uncertainty in the estimated concentrations. In this case, wavelength selection is important to reduce the collinearity before (or simultaneously) to the selection of the calibration samples. PCR does not present these collinearity problems since the scores are orthogonal.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

2.6 Collinearity in multivariate calibration

2.6.1 Definition of collinearity and singularity

Multiple linear regression regresses several independent variables x on a dependent variable y (eq 2.6). Collinearity (also called multicollinearity) is defined as approximate linear dependence of at least one of the columns $\mathbf{x}_j$ of $\mathbf{X}$ with other/s column/s[MLII]. Singularity accurs when the variables are perfectly correlated.

2.6.2 Problems caused by collinearity

Collinearity and singularity concern an ill-conditioning of the matrix $\mathbf{X}^T\mathbf{X}$ that causes numerical and statistical problems in MLR related to stability and ability for matrix inversion:

1. At least one diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$ is large and the associated least-squares estimated coefficient has large variance. The more collinear the x variables are, the more unstable becomes the linear system i.e. more susceptible to large changes in $\mathbf{b}$ produced by small changes in $\mathbf{X}$ or $\mathbf{y}$ due to noise. Unacceptable signs or too large values of the coefficients can be obtained [100] which affects their chemometric interpretation.

2. Numerical difficulties in $(\mathbf{X}^T\mathbf{X})^{-1}$, that cannot be calculated in case of singularity or may have large values in case of collinearity. For each exact linear dependence in the columns of $\mathbf{X}$ there is one zero eigenvalue of $\mathbf{X}^T\mathbf{X}$. Near-linear dependencies result in small eigenvalues.

3. Collinearity makes it more difficult to interpret the impact of each regressor on the response: correlated estimates cannot be interpreted separately and are unstable and unreliable. A regression coefficient is the partial derivative of the response with respect to a regressor variable[100]. It is desirable to have independent estimations of the coefficients.

4. The *t*-test can indicate statistical insignificance of the coefficients owing to large variances of regression coefficient estimates[76].

5. The prediction for new x measurements may be good at points with combinations of x's similar to those in the calibration data. Prediction at combinations different from these or extrapolation outside the range of the data can be adversely affected and have large errors.

6. Collinearity can exist in models with a good fit (a high multiple correlation coefficient). Since the residuals in the regression may be very small but the coefficients are estimated poorly, the traditional analysis of lack of fit does not signal potential collinearity problems[100].

2.6.3 A graphical representation of collinearity

The collinearity problem is illustrated in Figure 2.13, with a training set with two measured variables, x_1 and x_2. The *sample domain*[101] , rectangle ABCD, is delimited by the highest and lowest values of these variables. If the fitted model is the plane given by $y=b_0 + b_1x_1 + b_2x_2$, this can be interpreted as "a table whose legs are situated in the coordinates of each point and the lengths of the legs are the measured ys". Due to a different measurement error in each point, not all the "legs" fit the table so the inclination of the table (given by the model coefficients) has some uncertainty. This uncertainty is smaller in the direction AC since the table has legs at the extremes and an error in the length of a leg has less effect on the inclination of the table. The uncertainty is higher in the direction BD. A new prediction in a point in the direction AC has a small uncertainty since the model is stable in this direction, but a predicted point near the vertex B (the point in black) has a high uncertainty due to the uncertainty of the table in that direction.

The PCA decomposition used in the PCR model defines a new variable along the direction AC and another in the direction BD and the new experimental domain as the largest and smallest values of the scores along these two PCs. A sample in the vertex B would be detected as an outlier. In addition not considering the direction BC makes more stable predictions.

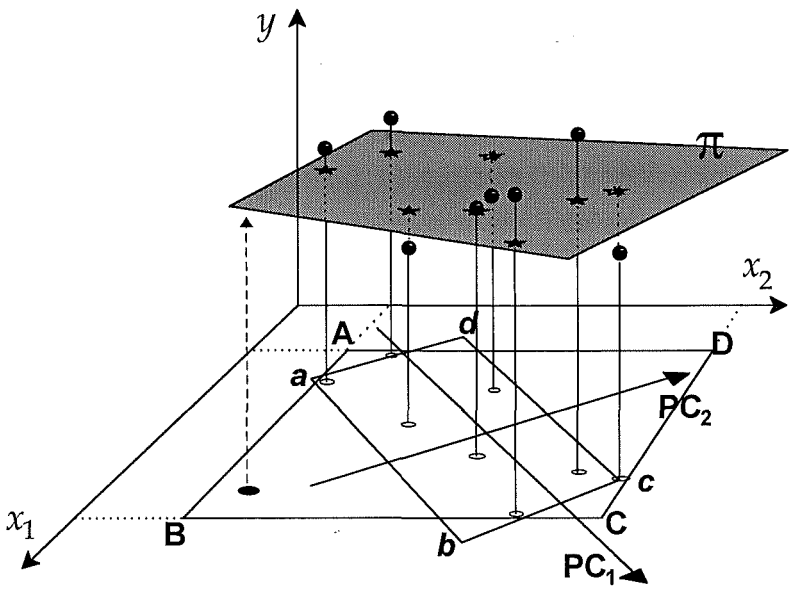


Figure 2.13. The training set has values that are very collinear: the x_1 values increase as the corresponding values for x_2 increase.

2.6.4 Detection and measures of collinearity

Several measures enable the extent of the collinearity problem to be evaluated: the correlation matrix of the regressor variables, the eigenvalues of the calibration matrix (and related measures such as the condition indices and the condition number), the tolerance and variance inflation factors of the estimated coefficients and the variance-decomposition proportions. A discussion about these measures can be found in references 34,100-106. The variance inflation factors (explained below) and the variance-proportion decompositions (explained in the section §4.3) are the diagnostic tools used in this thesis .

Variance inflation factors

The variance inflation factor (VIF) of the regression coefficient b_j is the jth diagonal element of the inverse of the correlation matrix of the variables. It can also be calculated as [106]

$$VIF_j = (1-R_j^2)^{-1} \quad (2.78)$$

where R^2_j is the multiple correlation coefficient of x_j regressed on all the other terms in the model. An equivalent definition is [10]:

$$VIF_j = UVIF_j \sum_{i=1}^{I} \left[x_{ij} - \bar{x}_j \right]^2 \quad (2.79)$$

The VIF_j measures the increase in variance in the fitted model due to the collinearity compared to a design of uncorrelated x-variables. The VIFj has a range 1 (non-correlated coefficients) to infinity (prefect correlation). Values larger than 1 indicate that the variable is affected by collinearity and larger than 10 that the correlation among the variables is so high that the coefficient is likely to be poorly estimated[107,108]. Since the maximum VIF is a lower bound on the condition number, a large VIF implies also a large condition number. The VIFs have been recommended as a general diagnostic measure of collinearity[104] and are used to measure if the selected points contain the sufficient information (if they are orthogonal enough) to estimate the model correctly [10,31]. Every wavelength or sample selection methodology can use the VIFs to measure the quality of the selected subset. In §4.6, the VIFs is shown to be equivalent to the Lorber's definition of selectivity. However, since the VIFs are a global measure, they do not indicate which regressors are involved and are unable to distinguish among coexisting spectral overlap situations of three or more components.

2.6.5 Influence of collinearity in multivariate calibration

Collinearity largely affects the prediction error in multivariate calibration models. Elimination of collinearity is important in the instrumental methods of analysis. Different sources of collinearity and their solutions are:

1. A bad experimental design of the analyte concentrations in the calibration samples. An example is the artificial calibration samples made by dilutions of a

single mixture with of all constituents of interest. The concentrations of all constituents vary together and their spectra all increase and decrease in sympathy. Multivariate models, which correlate changes in the concentrations to changes in the spectra, will fail since they will detect only one cause of variation regardless of how many constituents were mixed together in the original mixture. To an eigenvector-based model, one only factor will contain nearly all the variance in the data. A sample that does not have exactly the same ratio of constituent concentrations as the calibration samples will be predicted wrong. Collinearity can be reduced here with properly designed calibration mixtures having different ratios of the concentration on the components of interest.

2. Physical constraints in model or in data so that only certain combinations of the independent variables can be evaluated. This could be solved by adequately selecting the regression method so that it can handle collinear data, such as factor-based regression or ridge regression.

3. In CLS (eq 2.22), the overlap of the pure component spectra produces collinear columns in **S** and large variances and covariances of the concentration estimates. This gives an unstable system of equations and small relative changes in **r** due to measurement error can produce large relative changes in **c** so that misleading results can be obtained (e.g. negative concentration values for analytes that are present). The effects of collinearity can be reduced (but not eliminated unless completely selective sensors are available) with a correct choice of the wavelengths, which affects both the trueness and precision of the predicted concentrations. Criteria for wavelength selection are usually based on some measure of orthogonality in the **S** matrix, such as the selectivity by Lorber[17]. This and other criteria are discussed in the chapter §4.

4. In ILS, singularity is produced by a number of calibration samples inferior to the number of measured variables (over-estimated regression).This can be solved by using more calibration samples, removing redundant variables (e.g. with genetic algorithms or stepwise multiple linear regression, although this last method is not sensitive to the collinearity of the independent variables) or using factor-based models such as PCR and PLS, which reduce the number of regressor variables. Another source of collinearity in ILS is the correlation between the

columns of **R** produced, specially in NIR or UV-vis spectra where the number of responses can easily approach 1000, by the similar absorbances at adjacent wavelengths that tend to increase and decrease together in the calibration samples. This causes a large sensitivity of the estimated $\mathbf{b}_k$ to small changes in $\mathbf{c}_k$ (see §2.6.2). The large variance (low precision) of the coefficients produces large variances for the predicted concentration in unknown samples. For these reasons ILS is always accompanied of wavelength selection to reduce the collinearity. PCR or PLS are usually employed instead of ILS. These methods reduce the variance of the coefficients compared with ILS by discarding part of the information of the data to estimate the regression coefficients. The resulting estimators are biased, but they may be preferable to ILS.

ISBN:978-84-691-1875-7/DL. T-337-2008

2.7 References

1. Lorber A., Kowalski B.R. *J. Chemom.* 2 (1988) 93-109.
2. Searle S.R. *Matrix Algebra Useful for Statistics.* John Wiley & Sons. New York. 1982.
3. Manne R. *Chem. Intell. Lab. Syst.* 2 (1987) 187-197.
4. De Jong S. *J. Chemom.* 9 (1995) 323-326.
5. Deming S.N., Morgan S.L. *Experimental Design: a chemometric approach.* Elsevier. Amsterdam. 1987
6. Wold S., Sjöström M., Carlson R., Torbjörn L., Hellberg S., SkagerbergB., Wikström C., Öhman J. *Anal. Chim. Acta* 191 (1986) 17-32.
7. Hellberg S., Sjöström M., Skagerberg B., Wold S. *J. Med. Chem.* 30 (1987) 1126-1135.
8. Hellberg S., Sjöström M., Skagerberg B., Wikström, C., Wold S. *Acta Pharm. Jugosl.* 37 (1987) 53-65.
9. Peissik A. *Methodologie de la recherche experimentale: properties et caracteristiques des matrices d'experiences pour les modeles polynomiaux du second degre.* Thesis. 1995. Université d'Aix Marseille III.
10. Sergent M., Mathieu D., Phan-Tan-Luu R., Drava G., *Chem. Intell. Lab. Syst.* 27 (1995) 153-162.
11. Meloun M., Militký J., Forina M. *Chemometrics for analytical chemistry. Vol 2. PC-aided Regression and Related Methods.* Ellis Horwood: Great Britain 1994
12. Lebart L., Morineau A., Fénelon J.P. *Traitement des données statistiques. Méthodes et programmes.* Dunod. 2 Ed. Bordas. Paris 1982.
13. Weisberg S. *Applied Linear Regression* 2nd Edition, Wiley, New York 1985.
14. Atkinson A.C., Humter W.G. *Technometrics* 10 (1968) 271-289.
15. Atkinson A.C., Donev.A.N. *Optimum Experimental Designs.* Oxford Statistical Science Publications: Oxford 1992
16. Carlson R. *Design and optimization in organic synthesis.* Elsevier: The Netherlands 1992
17. Lorber A. *Anal. Chem.* 58 (1986) 1167-1172.
18. Goupy J. *Chem. Intell. Lab. Syst.* 33 (1996) 3-16.
19. Araujo P.W., Brereton R.G. *Trends Anal. Chem.* 15 (1996) 26-31.
20. Araujo P.W., Brereton R.G. *Trends Anal. Chem.* 15 (1996) 63-70.
21. Box G.E.P., Hunter W.G., Hunter J.S. *Statistics for Experimenters* Wiley: New York. 1978

22. Broudiscou A., Leardi R., Phan-Tan-Luu R., *Chem. Intell. Lab. Syst.*, 35 (1996) 105-116.
23. Pukelsheim F. *Optimal Design of Experiments*. Wiley. New York. 1993
24. Fedorov V.V. *Theory of Optimal Experiments*.(translated and edited by W.J. Studden and E.M. Klimko) Academic Press: New York, 1972
25. Kiefer J. *J. Royal. Statist. Soc. B.* 21 (1959) 272-319.
26. Pázman A. *Foundations of Optimum Experimental Design*. D.Riedel Publishing Company. Dordrecht. 1986.
27. Nishii R. *Discrete Mathematics*, 116 (1993) 209-225.
28. Pukelsheim F., Rosendberg J.L. *J. Amer. Stat. Assoc.* 88 (1993) 642-649.
29. Atkinson A.C. *Chem. Intell. Lab. Syst.*, 28 (1995) 35-47.
30. Khuri A.I., Cornell J.A *Response Surfaces. Designs and Analyses*. Marcel Dekker: USA 1987
31. Mathieu D. *Contribution de la Méthodologie de la Recherche Expérimentale à l'étude des relations Structure-Activité*. Thesis. Marseille, 1981.
32. Steinberg D.V., Hunter W.G. *Technometrics*, 26 (1984) 71-130.
33. Papakyriazis P.A. *J. Econometrics* 7 (1978) 351-372.
34. Willan A.R., Watts D.G.. *Technometrics* 20 (1978) 407-412.
35. De Aguiar P.F., Bourguignon B., Khots M.S., Massart D.L., Phan-Tan-Luu R. *Chem. Intell. Lab. Syst.*, 30 (1995) 199-210.
36. Franquart P. *Optimisations Multi-criteres et Methodologie de la Recherche Experimentale*. Thesis. Université d'Aix-Marseille III. 1992.
37. Naes T. *J. Chemom.* 1 (1987)121-134.
38. Kennard R.W., Stone L.A. *Technometrics* 11 (1969) 137- 148.
39. Jouan-Rimbaud D., Khots M.S., Massart D.L., Last I.R., Prebble K.A. *Anal. Chim. Acta* 315 (1995) 257-266.
40. Zemroch P. J. *Technometrics* 28 (1986) 39-49.
41. Martens H.; Naes T. *Multivariate Calibration*, Wiley: New York, 1989.
42. Sanchez E., Kowalski B.R. *J. Chemom.* 2 (1988) 247-263.
43. Kowalski B.R., Seasholtz M.B. *J. Chemom.* 5 (1991)129-145.
44. Thomas E.V. *Anal. Chem.* 66 (1994) 795A-804A.
45. Geladi P., Kowalski B.R. *Anal. Chim. Acta* 185 (1986) 1-17.
46. Martens H., Karstang T., Naes T. *J. Chemom.* 1 (1987) 201-219.
47. Marbach R., Heise H.M. *Chem. Intell. Lab. Syst.* 9 (1990) 45-63.
48. Frank I.E., Friedman J.H. *Technometrics* 35 (1993) 109-139.

49. Marbach R., Heise H.M. *Trends Anal. Chem.* 11 (1992) 270-275.
50. Brown C.W., Lynch P.F., Obremski R.J , Lavery D.S. *Anal. Chem.* 54 (1982) 1472-1479.
51. Haaland D.M., Thomas E.V. *Anal. Chem.* 60 (1988) 1193-1202.
52. Booksh K.S., Kowalski B.R *Anal. Chem. 66* (1994) 782A-804A.
53. Beebe K.R., Kowalski B.R. *Anal. Chem.* 59 (1987) 1007A-1017A.
54. Lang P.M., Kalivas J.H. *J. Chemom.* 7 (1993) 153-164.
55. Kalivas J.H., Lang P. M. *Chem. Intell. Lab. Syst.* 32 (1996) 135-149.
56. Maris M.A., Brown C.W., Lavery D.S. *Anal. Chem.* 55 (1983) 1694-1703.
57. Lorber A., Wagen L.E., Kowalski B.R., *J. Chemom.* 1 (1987) 19-31.
58. Faber K., Kowalski B.R. *J. Chemom.* 11 (1997) 181-238.
59. Bauer G., Wegscheider W., Ortner H.M. *Spectrochimica Acta 46B* (1991) 1185-1196
60. Otto M., Wegscheider W. *Anal. Chem.* 57 (1985) 63-69.
61. Schmidt P.C., Glombitza B.W. *Trends Anal. Chem.* 14 (1995) 45-48.
62. Naes T., Martens H. *J. Chemom.* 2 (1988) 155-167.
63. Kvalheim O.M. *Chem. Intell. Lab. Syst.* 8 (1990) 59-67.
64. Sekulic S., Seasholtz M.B., Wang Z., Kowalski B.R. *Anal. Chem.* 65 (1993) 835A-845A.
65. Lorber A., Kowalski B.R. *Appl. Spectrosc.* 42 (1988) 1572-1574.
66. Naes T. *Chem. Intell. Lab. Syst.* 5 (1989) 155-168.
67. Seasholtz M.B., Kowalski B.R. *J. Chemom.* 6 (1992) 103-111.
68. Lorber A., Harel A., Goldbart Z., Brenner. B. *Anal. Chem.* 59 (1987) 1260-1266.
69. Lorber A., Faber K., Kowalski B.R. . *Anal. Chem.* 69 (1997) 1620-1626.
70. Lorber A. , Kowalski B.R. *J. Chemom.* 2 (1988) 67-79.
71. Jouan-Rimbaud D., Massart D.L., Leardi R., De Noord O.E. *Anal. Chem.* 67 (1995) 4295-4301.
72. Leardi R. *J. Chemom.* 8 (1994) 65-79.
73. Vigneau E., Devaux M.F., Qannari E.M., Robert P. *J. Chemom.* 11 (1997) 239-249.
74. Shaffer R.E. Small, G.W. Arnold M.A. *Anal. Chem.* 68 (1996) 2663-2675.
75. Lorber A., Kowalski B.R. *Appl. Spectrosc.* 44 (1990) 1464-1470.
76. Militký, J., Meloun M. *Anal. Chim. Acta* 277 (1993) 267-271.
77. Osten D.W. *J. Chemom.* 2 (1988) 39-48.
78. Wold S., *Technometrics* 20 (1978) 397-405.
79. Wold S., Esbensen K., Geladi P. *Chem. Intell. Lab. Syst.* 2 (1987) 37-52.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

80. Kvalheim O.M., Karstang T.V. *Chem. Intell. Lab. Syst.* 7 (1989) 39-51.
81. Malinowski, E. R. *Factor Analysis in Chemistry*, 2nd Ed. John Wiley & Sons Inc: New York 1991.
82. Eastment H.T., Krzanowski W.J. *Technometrics* 24 (1982) 73-77.
83. Todeschini R. *Anal. Chim. Acta* 348 (1997) 419-430.
84. Sun J. *J. Chemom.* 9 (1995) 21-29.
85. Sun J. *J. Chemom.* 10 (1996) 1-9.
86. Jouan-Rimbaud D., Walczak B., Massart D.L., Last I.R., Prebble K.A. *Anal. Chim. Acta* 304 (1995) 285-295.
87. Sutter J.M., Kalivas J.H., Lang P.M. *J. Chemom.* 6 (1992) 217-225.
88. Xie Y., Kalivas J.H. *Anal. Chim. Acta* 348 (1997) 19-27.
89. Xie Y., Kalivas J.H. *Anal. Chim. Acta* 348 (1997) 29-38.
90. Höskuldsson A. *J. Chemom.* 6 (1995) 91-123.
91. Höskuldsson A. *J. Chemom.* 2 (1988) 211-228.
92. De Jong S., Ter Braak C.J.F. *J. Chemom.* 8 (1994) 169-174.
93. Geladi P. *J. Chemom.* 2 (1988) 231-246.
94. Lindgren F., Geladi P., Wold S. *J. Chemom.* 8 (1994) 377-389.
95. Lindgren F., Geladi P., Wold S. *J. Chemom.* 7 (1993) 45-59.
96. Marengo E., Todeschini R. *Chem. Intel. Lab. Syst.* 12 (1991) 117-120.
97. Rännar S., Lindgren F., Geladi P., Wold S. *J. Chemom.* 8 (1994) 111-125.
98. Karstang T.V., Toft J., Kvalheim O.M. *J. Chemom.* 6 (1992) 177-188.
99. Araujo P.W., Brereton R.G. *Trends Anal. Chem.* 15 (1996) 156-163.
100. Myers R.H *Classical and modern regression with applications 2nd edition.* Duxbury Press: Belmont 1990.
101. Mandel J. *J. Res. of the National Bureau of Standards* 90 (1985) 465-476.
102. Belsley D.A., Kuh E., Welsch R.E.. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* Wiley: New York 1980 Chapter 3.
103. Berk K.N *J. Amer. Statist. Assoc.* 72 (1977) 863-866.
104. H. M. Wadswoth. *Handbook of Statistical Methods for Engineers and Scientist.* McGraw Hill: USA 1990.
105. Snee R.D. *J. Qual. Tech.* 5 (1973) 67-79.
106. Kalivas J.H. *J. Chemom.* 3 (1989) 409-418.
107. Snee R.D. *Technometrics* 19 (1977) 415-428.
108. Marquardt D.W. *Technometrics.* 12 (1970) 591-612.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

Chapter 3

Selection of Calibration Samples and Factors in Principal Component Regression

ISBN:978-84-691-1875-7/DL: T-337-2008

3.1 Introduction

3.1.1 Aim of the chapter

The aim of this chapter is to propose a new methodology for selecting the best subset of calibration samples for PCR from the instrumental responses of a large set of samples. Only the selected samples are submitted to chemical analysis and calibration, thus reducing the time and cost of the calibration step. As a part of the methodology, a fast method for selecting the most important principal components for regression is developed.

3.1.2 Structure of the chapter

After the introduction, containing the aim of the chapter, its structure and a bibliographic revision, the sections §3.2 to §3.7 contain the main contents structured in different papers:

§3.2 is the paper *Selection of best calibration sample subset for multivariate regression.* Joan Ferré, F. Xavier Rius *Anal. Chem.* 68, (1996) 1565-1571. Here a sample selection methodology in PCR, based on the D-optimality criterion and the Fedorov's exchange algorithm, is described. The algorithm is also reviewed in the last part of this paper.

§3.3 is the paper *Determination of ethylene content in poly(propylene-ethylene) copolymers using near-infrared spectra (NIR) and multivariate calibration* Villagrasa C., Ferré J., Larrechi M.S., Rius F.X., García C. (*in preparation*). Here near-infrared (NIR) data of copolymers is used to compare the predictive ability of PLS and of PCR with the factors selected according to the methodology described in §3.2.

§3.4 is the paper *Constructing D-optimal designs from a list of candidate samples.* Joan Ferré, F. Xavier Rius *Trends Anal. Chem.* 16, (1997) 70-73. Here, the Fedorov's algorithm, that is seldom used in the analytical literature, is presented and compared

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

with the popular Kennard-Stone's algorithm[1] and the random division of samples into calibration and validation sets. Kennard-Stone's algorithm selects samples spread over the experimental domain. In contrast, the samples selected with the Fedorov's algorithm tend to lie at the extremes of the experimental domain for a linear model of first degree. These methods are applied in a MLR model to predict the octane index in fuel samples

§3.5 is the paper *Selection of calibration points for PCR in QSAR studies* . Joan Ferré, F. Xavier (*in preparation*) where the Fedorov's algorithm is applied in quantitative structure-activity relationship (QSAR) studies to select calibration points characterized by the scores on some PCs of a series of properties. The algorithm is an alternative to selecting the samples for its similarity to the points of a given experimental design (usually a factorial design).

§3.6 contains the paper *Assessing the validity of principal component regression models in different analytical conditions*. Rius A.; Callao M.P., Ferré J.; Rius F.X., *Anal. Chim. Acta* 337 (1997) 287-296. This paper proposes a methodology for assessing, before using the piecewise direct standardization (PDS) technique, if a PCR model is valid when the actual working conditions are different from those used for modeling. The contribution to this work consists on applying the D-optimality criterion for selecting, from a large set, the minimum number of samples that must be measured. These already analyzed samples can be used in the standardization process in case that it is necessary.

3.1.3 Bibliographic revision and comments

To estimate the coefficients in a multivariate calibration model a set of calibration samples with known instrumental responses (e.g. absorbance at different wavelengths) and analyte concentrations is required.

The design or selection of the calibration samples must consider and satisfy statistical and economical requirements. The economical aspect mainly comprises the cost and time involved to obtain the calibration samples. Measuring the instrumental

responses is supposed to be cheaper and less time-consuming than determining the analyte concentration. This last step may require cumbersome analyses with a well established or reference method of analysis, which might be expensive, slow or undesirable. The statistical performance characteristics are related to the quality of the values predicted with the model. The spectra and composition of the calibration samples should emulate the unknown samples as closely as possible. The composition should span the expected range of concentration values of the future unknown samples and the spectra should be representative of all the constituents that contribute to the instrumental response in the unknown sample to enable the model to recognize the information for the constituents of interest. All phenomena (with chemical, physical or other basis) that vary in the unknown samples and influence the instrumental measurements must also vary in the calibration set over the same ranges. Martens and Naes[2] and Gemperline[3] also commented these ideas.

Commonly, calibration sets have a relatively large number of samples (maybe hundreds) to achieve a statistical representation of all sample properties. Lorber and Kowalski[4] proved that, to improve the prediction quality, it is always advantageous to add samples to the calibration. Although the mathematical expression of the Sherman-Morrison-Woodbury theorem used in their proof contains a small erratum ($\mathbf{X}^T\mathbf{X}$ should be written instead of $\mathbf{X}$; the correct expression can be found in Meyers[5] page 459 or in Weisberg[6] page 293), the conclusions are not affected. However, due to the effort of analyzing each calibration sample with the well-established technique, the considerable cost of obtaining a large calibration set may not always be compensated by an equal increase in the quality of the model. Faber and Kowalski[7] commented that increasing the number of calibration samples above some limit has only a marginal effect on the prediction error. In addition, Honigs *et al.*[8] mentioned some specific drawbacks to the use of a large sample set, such as the possibility that a property that is only present in a few calibration samples be ignored by the model if many other samples do not present this property. A small training set selected to contain a high degree of variability could avoid that problem.

The experimenter usually employs his/her subjective criterion to decide when the number of the calibration samples is 'sufficient' and when their composition spans correctly the experimental domain. The number considered as 'sufficient' depends on the nature, cost and difficulty in obtaining the samples. Moreover, the available

samples are usually randomly divided into calibration and validation sets. Several authors[9,10] found that calibrations based on random choice may perform well but there is also a chance to obtain much worse results than with a careful selection using a methodological approach. The validation of the models also depends on how well the calibration set represents the validation set.

Owing to the large importance of the calibration set on the predictive ability of the model, the selection of the calibration samples should not depend upon scarcely rigorous criteria. The idea here is to employ a mathematical criterion to reduce the cost of the calibration by selecting an adequate number of calibration samples that gives a compromise between the performance criteria and the cost of the model. Since it is easy to perform measurements on a large number of samples, there is a large probability that the calibration set contains the most important variations in the data.

3.1.3.1 Bibliographic revision of calibration sample selection

Table 3.1 resumes different approaches found in the literature for selecting calibration samples in PCR from a large list when they cannot be synthesized. These procedures often use spectral data, which can be collected with little labor and cost, and characterize the samples by their scores on a certain number of principal components (PCs). Other approaches found, although not based on PCR, are also indicated but not commented.

3.1.3.2 Comments to the existing approaches

The idea of all these methods is to select a representative sample among other similar and avoid discarding useful information. The approaches based on the Kennard-Stone-like algorithms[1,11] and clustering[10] have the interesting advantage of collecting points evenly spread over the whole experimental domain and span the variation as uniformly as possible. In this way, the selected samples are not too close to any of the others and can be used for check the fit of the model or add new terms if necessary. However, some general objections can be made to the methods indicated in the Table 3.1.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

Table 3.1. Sample selection approaches proposed by several authors.

Authors	*Proposed approach. Comments*
Hruschka and Norris[12]	Selection using concentration and instrumental responses of all the samples in ILS.
Honigs *et al.*[8]	Subtractions from the spectral data to choose the spectrally unique samples for calibration from a large set. The algorithm spans the spectral variation as much as possible.
Zemroch[13]	Clustering of the candidate points and selection of one point from each cluster in MLR.
Naes[10]	Clustering of the samples using their scores on a certain number of PCs of NIR spectra. The sample farthest away from the center of every cluster is selected as representative.
Puchwein[9]	Iterative elimination of similar samples from a large data set using sample scores and distances between data points. The samples retained for calibration have the largest Mahalanobis distance from the origin and are representative of the complete original data set.
Lorber and Kowalski[4]	Algorithm for sensor selection that can be modified for sample selection. The calibration samples, that are optimal for all analytes, are selected after measuring the response of the unknown sample. The optimal calibration set may change (thus different samples must be analyzed) for each unknown sample., which does not reduce the cost of the calibration.
Schostack and Malinowski[14]	Iterative key set factor analysis (IKSFA) to select the key set, the preferred set of analytical wavelengths or calibration samples that best characterizes a multicomponent system.
Isaksoon and Naes[15]	Compared Naes[10] and Honigs *et al.*[8] approaches in PCR. The Naes[10] approach performed better in terms of prediction error.
Kalivas[16]	Generalized simulated annealing (GSA) to select calibration samples from a set of NIR spectra by minimizing the Mahalanobis distance between the unknown sample and the average spectrum. Not applied to PCR but to PLS with one latent variable. The reason for this one latent variable was not indicated, nor the criteria to estimate the appropriate number of calibration samples.
Hitchcock *et al.*[17]	Design of optimal calibration concentration matrices for spectroscopic data and PLS. They assume that the analyst knows the components present in the unknown sample, the proper number of calibration samples and is able to artificially generate them.
Tong-Hua *et al.* [18]	A genetic algorithm is used to select calibration samples. Its number is decided a priori.
Marengo and Todeschini[11]	Algorithm for selecting experiments uniformly distributed from the set of candidates using the original variables, not PCA scores. It does not require any preliminary hypothesis about a regression model. Similar to Kennard-Stone's algorithm[1].
Aastveit and Marum[19]	They compared different strategies for sample selection in PCR, including the Naes's[10] clustering method. Local calibration methods performed the best.
Naes and Isaksoon[20]; Araujo and Brereton[21]	General rules for selection of samples for calibration, specially for ILS models
Jouan-Rimbaud *et al.* [22]	Kennard-Stone algorithm to select calibration samples in ILS.

One objection is the lack of a clear mathematical criterion to decide the sufficient (or optimal) number of calibration samples to correctly estimate the model coefficients. The authors either did not give any criterion[4,11,16] or decided this number beforehand depending on how many samples one wants to or can afford to use in the calibration [8,10,13,]. For example, Naes[10] divided randomly the available samples into calibration and validation sets, and decided arbitrarily a number of samples from the calibration set to be selected using clustering. Honigs *et al.*[8] continued the iterative procedure until the desired number of spectra was selected. Puchwein[9] submitted to factor analysis the raw data of the reduced number of samples and the transformation matrix derived was used to recalculate the factor scores of the whole data set. The subset was assumed to still represent the original set if the redefined subset region contained all or at least most of the original samples. However, he was not able to formulate a rule to stop the sample reduction automatically when the minimum subset was reached. When the experimenter is not able to decide the exact number of samples that are adequate to assure the quality of the estimated model coefficients, these approaches may fail.

Another limitation is that the selections based on distances require the PCs used in the model (which should be the ones with the best predictive ability) to be specified beforehand (e.g. the scores on the important PCs are used to compute the Mahalanobis distance in the clustering method[10]). However, this cannot be known only from the instrumental responses matrix (see §2.4.2.4.3); they must be selected considering the predictive ability of PCR models made with an increasing number of factors. Moreover, the factors should be included in these models in order of their correlation with the concentration, not in order of the percentage of explained variance of the data matrix. To make sure that the optimal number of components is incorporated it was suggested using experience with similar systems on how many PCs had main predictive relevance or to randomly select a few calibration samples and estimate the optimal number of factors with cross-validation or leverage correction [10,15]. This number could then be used in a search for additional calibration samples according to the selection procedure. Naes[10] selected the optimal factors for clustering using all the available samples and their analyte concentrations. So did Xie and Kalivas[23] and Sutter *et al.*[24] and in their optimization procedure to find the optimal set of factors. These approaches cannot be used here since the aim of the methodology is avoid analyzing all the samples. Puchwein[9] first considered the factors required to regenerate the raw data matrix within the measurement error and

the factors were selected by regressing the concentration against the scores of each factor individually. The factors were introduced into the final model by the absolute value of their regression coefficients in decreasing order. Isaksoon and Naes[15] selected the PCs with the largest variance but did not give any criterion to justify the number of factors used. They indicated that further studies on how to select an optimal number of PCs should be performed.

Considering the mentioned comments it can be stated that a strategy of selection of the calibration sample subset for PCR requires:

1. A criterion for judging the quality of each subset of N candidates. The selected subset is the one that optimizes this criterion over all the other possible subsets. Faber and Kowalski[7] also indicated that the a suitable selection criterion should provide a design of the calibration samples good enough to avoid extrapolations.

2. An optimization algorithm to find the optimal subset avoiding the examination of all possible combinations of subsets of samples.

3. A criterion for comparing the subsets with a different number of samples, so that the optimal N can be decided.

4. A method for selection of the best predictive factors for PCR in case that the optimization criterion needs them to be specified beforehand

A selection methodology should consider the above indicated steps using the instrumental responses of a large set of samples but analyzing only the minimum number required. This has been solved in the paper presented in §3.2.

3.1.4 References

1. Kennard, R.W. ; Stone, L.A. *Technometrics* 11 (1969) 137- 148.
2. Martens H.; Naes T. *Multivariate Calibration*, Wiley: New York, 1989.
3. Gemperline P.J. *J. Chemom.* 3 (1989) 549-568.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

4. Lorber A. , Kowalski B.R. *J. Chemom.* 2 (1988) 67-79.
5. Myers R.H *Classical and modern regression with applications 2nd edition.* Duxbury Press: Belmont 1990.
6. Weisberg S. *Applied Linear Regression* 2nd Edition, Wiley, New York 1985.
7. Faber K., Kowalski B.R. *J. Chemom.* 11 (1997) 181-238.
8. Honigs D.E.; Hietfje G.M.; Mark H.L.; Hirschfeld T.B. *Anal. Chem.* 57 (1985) 2299-2303.
9. Puchwein G. *Anal. Chem.* 60 (1988) 569-573.
10. Naes, T. *J. Chemom.* 1 (1987)121-134..
11. Marengo E. ; Todeschini R. *Chem. Intel. Lab. Syst.* 16 (1992) 37-44.
12. Hruschka W.R., Norris K.H. *Applied Spectrosc.* 36 (1982) 261-265.
13. Zemroch, Peter J. *Technometrics* 28 (1986) 39-49.
14. Schostack, K. J.; Malinowski, E. R. *Chem. Intel. Lab. Syst.* 6 (1989) 21-29.
15. Isaksoon T.; Naes T. *Appl. Spectrosc.* 44 (1990) 1152-1158.
16. Kalivas, J.H. *J. Chemom.* 5 (1991) 37-48.
17. Hitchcock K., Kalivas, J.H., Sutter J.M. . *J. Chemom.* 6 (1992) 85-96.
18. Tong-Hua Li, Lucasius C.B., Kateman G. *Anal. Chim. Acta* 268 (1992) 123-134.
19. Aastveit A.H.; Marum P., *Appl. Spec.* 47 (1993) 463-469.
20. Naes T., Isaksoon T. *NIR news* 5 (1994) 16-17.
21. Araujo P.W., Brereton R.G. *Trends Anal. Chem.* 15 (1996) 156-163.
22. Jouan-Rimbaud D., Khots M.S., Massart D.L., Last I.R., Prebble K.A. *Anal. Chim. Acta* 315 (1995) 257-266.
23. Xie Y., Kalivas J.H. *Anal. Chim. Acta* 348 (1997) 19-27.
24. Sutter, J.M.; Kalivas J.H. ; Lang P.M. *J. Chemom.* 6 (1992) 217-225.

3.2 Selection of the Best Calibration Sample Subset for Multivariate Regression

Anal. Chem. 68 (1996) 1565-1571

Joan Ferré and F. Xavier Rius.

Departament de Química. Universitat Rovira i Virgili.
Pl. Imperial Tarraco, 1, 43005-Tarragona. SPAIN

This paper discusses a methodology for selecting the minimum number of calibration samples in principal component regression (PCR) analysis. The method uses only the instrumental responses of a large set of samples to select the optimal subset, which is then submitted to chemical analysis and calibration. The subset is selected to provide a low variance of the regression coefficients. The methodology has been applied to UV-visible spectroscopy data to determine Ca^{2+} in water and near-IR spectroscopy data to determine moisture in corn. In both cases, the regression models developed with a reduced number of samples provided accurate results. As far as precision is concerned, a similar root-mean-squared error of cross-validation (RMSECV) is found when comparing the new methodology with the results of the regression models that use the complete set of calibration samples and PCR. The number of analyzed samples in the calibration set can be reduced by up to 50%, which represents a considerable reduction in costs.

Received for review May 22, 1995. Accepted January 18, 1996.

Multivariate regression methods [1-4] such as classical least-squares (CLS), inverse least-squares (ILS), principal component regression (PCR), or partial least-squares regression (PLSR) enable mathematical models to be developed that relate multivariate instrumental responses r_j (e.g.: spectral intensities) from many calibration samples to the known analyte concentrations in these samples (c_i) according to eq 1:

$$c_i = f(r_1, r_2, \ldots, r_j) + e_i \qquad (1)$$

where e_i is the residual associated with the ith concentration. This relationship can then be used to predict analyte concentrations for unknown samples from their instrumental responses. When calibration standards are not easily synthesized (e.g., natural samples), they are selected from among all the available samples. For this calibration set to be obtained, the instrumental responses must be measured (which is usually relatively quick and easy, as in spectroscopic analysis) and the analyte concentrations determined. These concentrations are usually determined with a reference or well-established method that may be slow, expensive, or cumbersome. Lorber and Kowalski[5] showed that, to establish suitable prediction models, the higher the number of calibration samples providing supplementary information the better. However, this might mean that the cost and time spent in obtaining a high number of calibration standards is not affordable in cost-effectiveness terms. The analyst might be interested in using not all but only the minimum (or a reduced) number of calibration samples, provided that the developed model is able to furnish prediction values of the desired quality.

In this paper we report a procedure for selecting an adequate subset of calibration samples for PCR from the instrumental responses of a large number of samples. Only the selected samples are submitted to the more time-consuming chemical analysis and to principal component modeling. The selected samples give the lowest variance for the estimated regression coefficients and enable the principal components (PCs) that provide the best predictive PCR model to be selected. The results compare well to the ones obtained using the complete set of calibration samples.

Several procedures for selecting a subset of calibration samples for PCR from a large data set have been proposed.[6,7] Clustering,[8] iterative elimination of similar samples by using the Mahalanobis distance,[9] and iterative key set factor analysis

(IKSFA)[10] have been applied to spectral data and make use of the scores on a certain number of PCs. In quantitative structure-activity relationship studies (QSAR), principal component analysis followed by sample selection to fit factorial and fractional factorial designs have been reported.[11-15] Other approaches for selecting calibration samples, although they do not focus on PCR, have been used and compared.[16-21] Recently, Naes and Isaksoon[22] gave some general principles for selecting calibration samples.

Although promising results have been obtained, some general objections can be made to these approaches. While they aim to span as much of the experimental domain as possible, the mathematical expression of the regression model is seldom considered. Moreover, no unambiguous mathematical criterion is reported (except for Puchwein's approach[9]) to decide how many calibration samples are required. The experimenter must usually decide a priori what a "sufficient" number of calibration standards is, using subjective criterion. In such cases, there is no guarantee that the selected samples will contain the necessary information to build a model which should provide accurate and precise predictions throughout the experimental domain. Moreover, as is shown below, the random separation of samples into calibration and test sets can give models with poor prediction ability if the calibration set does not contain enough information to allow correct estimation of the regression coefficients.

In the field of experimental design theory,[23-29] a variety of algorithms[25-29] have been used to select subsets of calibration samples for multiple linear regression (MLR) models. A usual selection criterion is to minimize the variance of the estimated regression coefficients, but this is very time-consuming when the number of predictor variables is high and is not suitable for highly collinear data. These two characteristics are common, for example, in spectroscopic data. PCR can overcome these problems since it can deal with collinear data and a large number of variables.

The methodology presented here uses Fedorov's[28-29] exchange algorithm to select an appropriate set of calibration samples for PCR models after scaling the scores. This algorithm makes use of the mathematical expression of the model, so the experimenter must know which PCs are relevant for regression in order to include them in the model. As this is a rather strict condition when the experimenter faces the regression problem for the first time, the PCs that are important for regression are

found by building a preliminary screening model that uses the minimum number of necessary samples. The important factors are selected from the absolute values of this model's coefficients. Subsequently, a definitive model containing the selected factors can be postulated, and the definitive sample subset that will be used to build it is selected. Finally, the model is validated by using the cross-validation technique.

Background and Theory

Notation. Matrices are represented by bold capital letters, column vectors by bold lowercase letters, and scalars by italic characters. The superscript T means transposed. The subindices in a matrix indicate its dimensions. Let $\mathbf{R}_{I\times J}$ be the column mean-centered matrix of instrumental response data for I samples and J sensors and $\mathbf{c}_{I\times 1}$ the vector of the kth analyte concentration in the I calibration samples.

Principal component regression formulation. In PCR, $\mathbf{R}_{I\times J}$ is decomposed according to

$$\mathbf{R}_{I\times J} = \mathbf{T}_{I\times P}\mathbf{P}_{J\times P}^{T} + \mathbf{E}_{I\times J} \qquad (2)$$

where the columns in $\mathbf{T}_{I\times P}$ are $P \leq min(I,J)$ uncorrelated underlying factors that might be important for prediction, $\mathbf{P}_{J\times P}$ is the loading matrix, and $\mathbf{E}_{I\times J}$ is a matrix of residuals. After determining which $Q \leq P$ principal components are important for regression, $\mathbf{c}_{I\times 1}$ is regressed versus $\mathbf{T}_{I\times Q}$ according to

$$\mathbf{c}_{I\times 1} = \mathbf{T}_{I\times(Q+1)}\beta_{(Q+1)\times 1} + \mathbf{f}_{I\times 1} \qquad (3)$$

where $\beta_{(Q+1)\times 1}$ is the vector containing the regression coefficients, a column vector of ones has been appended to $\mathbf{T}_{I\times Q}$ to account for a constant term, and the elements of $\mathbf{f}_{I\times 1}$ are the calibration error terms. An estimate of β can be found using the least-squares method.[2]

The Q factors that provide the best predictive model are usually found by cross-validation[2] of several regression models built with a different number of factors. Since the factors associated with the largest eigenvalues do not necessarily give the best

ISBN:978-84-691-1875-7/DL: T-337-2008

predictive model,[30,31] all possible combinations of factors should be checked. Optimization methods such as generalized simulated annealing (GSA)[30] enable the best subset to be found and make it unnecessary for all possible combinations to be checked. Although GSA has provided promising results, it uses the whole set of samples and may often be quite time-consuming. We propose a faster approach based on screening the factors which are important for modeling the concentration using instrumental response and analyte concentration values for a reduced number of samples.

Sample Selection and Model Building. The devised procedure consists of the following steps:

(1) $\mathbf{R}_{I\times J}$ is first decomposed according to eq 2. The P factors in $\mathbf{T}_{I\times P}$ that contain important information are selected on the basis of previously acquired experience about similar systems, criteria about significant eigenvalues (e.g., Malinowski's IND function[32]), or the commonly used cross-validation technique.[33,34] It is preferable to overdetermine P to ensure that all factors containing information are taken into account.

(2) The range-midrange transformation[35] is performed on $\mathbf{T}_{I\times P}=\{t_{ip}\}$ for every sample $i=1, ..., I$ and every factor $p=1, ..., P$ so as to obtain $\mathbf{S}_{I\times P}=\{s_{ip}\}$:

$$s_{ip}=(t_{ip}-C_p)/R_p \qquad (4)$$

where:

$$C_p= [\max (t_{ip}) + \min (t_{ip})] / 2$$
$$R_p= [\max(t_{ip}) - \min (t_{ip})] / 2$$

with $\max(t_{ip})$ and $\min(t_{ip})$ being the highest and lowest score values for the pth factor. This scaling technique makes the scaled scores s_{ip} for every factor span the range [−1,+1]. Although this scaling technique can be sensitive to outliers, it is essential since it enables the regression coefficients calculated in step 4 to be compared.

(3) A subset of N samples ($P+1\leq N< I$) is selected according to the procedure developed in the next section and their analyte concentrations ($\mathbf{c}_{N\times 1}$) are chemically determined.

(4) The coefficients in model eq 5 are determined:

$$\mathbf{c}_{N\times 1} = \mathbf{S}_{N\times(P+1)}\beta_{(P+1)\times 1} + \mathbf{f}_{N\times 1} \quad (5)$$

where $\mathbf{S}_{N\times(P+1)}$ is a submatrix of $\mathbf{S}_{I\times P}$, to which the column vector of ones has been appended to account for the constant term and whose rows are the N samples selected in step 3. The magnitude of each estimated coefficient in eq 5 indicates the ability of its corresponding factor to model the concentration values. As the regressor variables are equally scaled between -1 and +1, the absolute values of the coefficients can be compared among one another. A high value indicates that the corresponding factor explains a considerable variance of the concentration. The Q factors with largest coefficients are selected for the final regression model since they provide the best modelling ability. The factors with coefficient values near to zero only model random error and are discarded. Should irrelevant factors be included in the model, the quality of the prediction will decrease.

To classify a factor as non-significant, two alternative approaches are presented here: a t-test for every coefficient[36] and the leave-one-out cross-validation error for a series of models made with the selected samples and by adding the factors in decreasing order according to the absolute value of their coefficients. The factors that give the model with the minimum error of prediction are selected.

(5) The final PCR model is built with the Q important factors and a new selected subset of M samples according to eq 6:

$$\mathbf{c}_{M\times 1} = \mathbf{S}_{M\times(Q+1)}\beta_{(Q+1)\times 1} + \mathbf{f}_{M\times 1} \quad (6)$$

where $\mathbf{S}_{M\times(Q+1)}$ is a submatrix of $\mathbf{S}_{I\times P}$ in which each row corresponds to a selected sample, each column corresponds to a selected factor, and a column of ones has been appended. The analyte concentrations of these new selected samples ($\mathbf{c}_{M\times 1}$) are determined according to the chemical procedure used. The final model can be validated using the cross-validation technique. Alternatively, the samples that have not been used in the model building step can be used as a test set to validate the model, but this requires analyzing these latter samples. The selection of a subset of test samples for model validation could be the subject for future research.

ISBN:978-84-691-1875-7/DL: T-337-2008

Selection of the Optimal Set of Calibration Samples. The criterion for sample selection is that the N calibration samples should provide regression coefficients with the lowest variance of all of the subsets of N samples. The global precision of the estimated coefficients in eq 5 is given by the volume of their 100(1-α)% confidence region, which is proportional to $[\mathrm{Det}(\mathbf{S}_{N\times(P+1)}{}^{\mathrm{T}}\mathbf{S}_{N\times(P+1)})]^{-1/2}$, where Det denotes determinant.[24,27-29] Maximizing $\mathrm{Det}(\mathbf{S}_{N\times(P+1)}{}^{\mathrm{T}}\mathbf{S}_{N\times(P+1)})$ by selecting which N samples are included in $\mathbf{S}_{N\times(P+1)}$ minimizes the volume of the confidence region and helps to achieve minimum variance in the coefficients[25]. This criterion is known as the D-optimality criterion. Every definite model, implicit in the matrix $\mathbf{S}_{N\times(P+1)}$ (and later in $\mathbf{S}_{M\times(Q+1)}$) in which each column is related to a coefficient, has a different optimal calibration set.

Although a search for all combinations of N samples ensures that the subset that maximizes $\mathrm{Det}(\mathbf{S}_{N\times(P+1)}{}^{\mathrm{T}}\,\mathbf{S}_{N\times(P+1)})$ is found, the time required for such a search makes it impractical when the number of available samples, I, is high. Several algorithms make examining all possible combinations unnecessary.[25,27] We used Fedorov's exchange algorithm[27-29] found in the NEMROD 3.0 software package,[37] since it is specially designed to search for D-optimal subsets from a list of I candidate samples. D-optimal subsets are found for different N: from N equal to the number of coefficients in the PCR model (i.e, the minimum required to solve the linear system of equations) to a number $N<I$ defined according to the user's needs. Of the D-optimal subsets, the one that is selected for calibration is the one that contains the maximum information *per sample* to estimate the coefficients, which is given by $\log(\mathrm{Det}(\mathbf{M}_N))$ with $\mathbf{M}_N=(\mathbf{S}_{N\times(P+1)}{}^{\mathrm{T}}\,\mathbf{S}_{N\times(P+1)})/N$.[29]

A similar procedure is used to select the final subset of M samples in eq 6, and the matrix $\mathbf{S}_{M\times(Q+1)}$ is used. As the optimal calibration subset is model-dependent, the N samples used to build eq 5 are not necessarily the most suitable for eq 6. Since the aim is to use the minimum number of calibration samples, rather than discard the previously analyzed N samples, the selection algorithm will add (if necessary) M-N samples to the already analyzed N samples so that the D-optimal subset contains the necessary information for a good estimation of the coefficients of the final model.

It should be noticed that the selection algorithm uses only the instrumental responses. Hence, a selected calibration set used in the screening step, $\mathbf{S}_{N\times(P+1)}$ is

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

optimal for all the analytes present in a sample, provided that the same model is postulated. This is no longer true for $\mathbf{S}_{M\times(Q+1)}$, which includes only the factors that are important for predicting each specific analyte.

Validation of the Methodology Developed. To assess the accuracy and precision of the PCR model built with a selected subset, the selected M samples are used as the calibration set while the remaining (I-M) samples are used as the test set. The accuracy is checked by a joint statistical test for the slope and the intercept of the linear regression between the measured versus predicted concentration values in the test set.[38] The multivariate model is regarded as being accurate if the theoretical values of intercept zero and slope unity are included within the ellipse which describes the joint confidence interval of the calculated straight line. The precision is measured by the root-mean-square error of prediction[2] (RMSEP) for the test set. In addition, the analyst can calculate the root-mean-square error of cross-validation (RMSECV) for the model built with the M selected samples, given in eq 7:

$$\text{RMSECV} = [\sum_{i=1}^{M} (c_{i_{\text{CV}}} - c_i)^2 / M]^{1/2} \qquad (7)$$

where $c_{i\text{CV}}$ is the predicted concentration for the ith sample in a model developed without the ith sample. The RMSEP of these models was also compared with the root mean square error of cross-validation (RMSECVT) for the model developed using the I available samples for calibration.

Experimental section

Samples and software. The following sample sets were used to check the validity of the method proposed for sample selection.

(1) Data set I consists of 24 UV–visible spectra. Rius et al.[39] determined Ca^{2+} by using the absorbance of their complexes with 2,2'-(1,8-dihydroxy-3,6-disulfonaphthylene-2,7-bisazo)-bis(benzenearsonic acid) (arsenazo III) in 24 natural water samples. Each spectrum consists of 101 variables, corresponding to the

absorbance values at wavelengths from 450nm to 650 nm. The actual content of calcium in water samples was determined by AAS and ranges between 3.1 and 40.6 ppm.

(2) Data set II consists of near-IR spectra of 46 corn samples at 19 fixed wavelengths reported by Puchwein.[9] The moisture of corn is the constituent of interest, and its content was determined by oven-drying. It ranges between 3.63% and 19.39%.

All computations were performed with home-made Matlab[40] subroutines. The instrumental response matrices were first mean-centered and Matlab SVD was used to evaluate the factors. The Matlab source codes are available on request.

Results and discussion

Data set I: Calcium in water samples. *Selection of the Important Factors To Be Used in the Screening Model.* P=10 factors were regarded as possibly containing important information according to PCA cross-validation. Hence, the minimum number of samples in the calibration set was 11 to enable the constant term and the regression coefficients associated to each factor in eq 5 to be estimated. Fedorov's exchange algorithm searched for the subset of N=11-24 samples that maximize $\mathrm{Det}(\mathbf{S}_{N\times 11}{}^{T}\mathbf{S}_{N\times 11})$. The plot $\log(\mathrm{Det}(\mathbf{M}_N))$ versus number of selected samples N (Figure 1) shows that the subset containing 15 samples ($\mathbf{S}_{15\times 11}$) has the maximum information per sample. Therefore, $\mathbf{S}_{15\times 11}$ and the calcium concentration for these 15 samples were used to build the screening regression model. The regression results are summarized in Table 1. Comparing the values in column 6 with the tabulated t-value for α =0.05 and 4 degrees of freedom (i.e, 15 samples minus 11 coefficients), $t_{0.05,4}$= 2.13, the factors numbered 7, 8 and 10 are discarded for regression. The same result is found by looking at the minimum cross-validation error.

Final regression model. The exchange algorithm was run again to maximize $\mathrm{Det}(\mathbf{S}_{M\times 8}{}^{T}\mathbf{S}_{M\times 8})$, where each column in $\mathbf{S}_{M\times 8}$ corresponds to the sample scores on the selected factors 1–6 and 9, and there is a column of ones to account for the constant

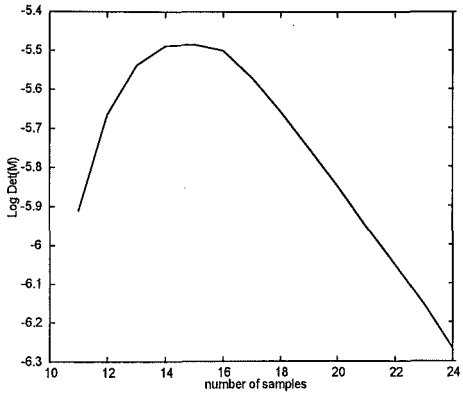


Figure 1. Number of selected samples (N) versus log(Det($\mathbf{M}_N$)). Calcium, 10 factor model.

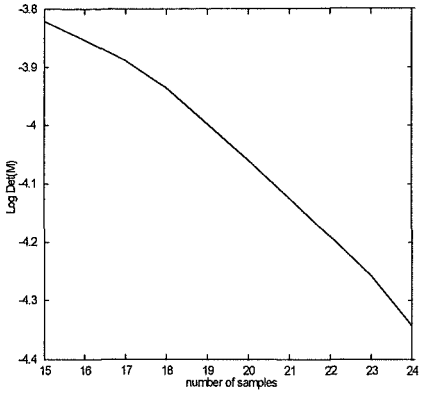


Figure 2. Number of selected samples (M) versus log(Det($\mathbf{M}_M$)). Calcium in water, seven selected factor model.

Table 1. Calcium in water samples. Regression Results for the Screening Model of 10 Factors and 15 Selected Samples [a]

PC	eigenvalue ($\times 10^{-3}$)	% variance	cumulative % variance	coeff	t-test	CV error
1	67.810	81.87	81.87	16.79	9.77	7.51
2	14.013	16.92	98.79	8.20	5.25	6.71
9	0.001	0.00	98.79	4.44	2.78	6.07
3	0.777	0.94	99.73	4.35	3.09	5.85
6	0.003	0.00	99.73	3.83	2.26	5.86
4	0.210	0.26	99.99	3.73	2.33	6.08
5	0.003	0.00	99.99	3.55	2.47	4.45
8	0.001	0.00	99.99	1.48	0.85	5.29
7	0.001	0.00	99.99	1.17	0.79	6.30
10	0.001	0.00	99.99	0.66	0.46	8.29

[a] Column 1 lists the PCs numbered according to decreasing eigenvalues. Columns 2-4 list the eigenvalue associated with each PC, the percentage of explained variance and the cumulative percentage of explained variance, respectively. Column 5 lists, in decreasing order of magnitude, the absolute value of the regression coefficient corresponding to each factor in column 1. Column 6 contains the calculated t-values (for comparison, tabulated t-value: $t_{0.05,4}$= 2.13). The last column shows the prediction error according to the leave-one-out cross-validation procedure for the selected samples using the factors cumulatively according to the ordered list in column 1.

term. To use the already-analyzed samples, the algorithm added samples to the 15 previously selected ones. The subset with maximum log(Det($\mathbf{M}_M$)) did correspond to the same 15 samples since any addition of new samples gave rise to a decrease in the log(Det($\mathbf{M}_M$)) (Figure 2). Thus, the 15 previously selected samples already contained enough information for the final model and no additional samples were needed, making any further analyte determination unnecessary.

Model validation. The results of the PCR model built with 15 samples and factors 1-6 and 9 are shown in Table 2. All the factors used in the final model are important, as shown by the *t*-test (tabulated $t_{0.05,7}$= 1.89) and the cross-validation error that reaches a minimum when the model is made with all the selected factors. The nine samples not used for model building were used as a test set. From the *F*-test for the joint confidence interval for the slope and intercept of the linear regression between the measured versus predicted concentration values, the model was considered to be accurate at an α = 0.425 level of significance. As far as precision was concerned, a very acceptable value of RMSEP = 1.86 was obtained. On the other hand, the value of RMSECV = 4.45 is higher than the RMSECVT = 3.00 obtained using the initial 24 samples. This could be explained if all 15 samples are important for building the PCR model. Deleting only one sample to calculate RMSECV using the leave-one-out procedure can considerably change the model, giving rise to a loss in precision.

Table 2. Calcium in Water Samples. Regression results for the Final Model with 15 Selected Samples and Factors 1-6 and 9[a]

PC	coeff	*t*-test	CV error
1	17.13	11.85	7.51
2	7.73	6.05	6.71
3	4.44	3.66	6.01
9	4.36	3.17	5.85
6	3.64	2.51	5.86
5	3.34	2.73	5.03
4	3.33	2.49	4.45

[a] The columns have the same meaning as columns 5-7 in Table 1. The first column is the number of each PC assigned in table 1, listed according to the absolute values of the coefficients for the final regression model.

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

Validation of the methodology. The performance of the developed methodology was also compared with the following methods for PCR modeling by evaluating their prediction ability. (a) The commonly used PCR method, where the complete set of samples is used for calibration and the factors are introduced into the model in decreasing order of eigenvalues. The results are given in Table 3, column A. (b) All possible models made with all possible combinations of PCs, where the complete set of samples is used for calibration and the model with minimum RMSECVT is selected for each number of factors. The results are given in Table 3, column B. (c) The same procedure as (b) but using only the 15 selected samples. The RMSECV results are listed in column C in Table 3.

Table 3, column B shows that models built with a subset of selected factors provide smaller RMSECVT values than models in which the factors are introduced in decreasing order of their eigenvalues (Table 3, column A). In this case, factor 7 probably models information not related to the calcium concentration, thus

Table 3. RMSECV Values of Regression Models for Calcium with a Different Number of Factors[a]

A) Complete set [a]	B) Complete set [b]	C) 15 selected samples [c]
15.39 (zero factors)	15.39 (zero factors)	18.51 (zero factors)
6.78 1	6.78 1	7.51 1
5.61 1 2	5.61 1 2	6.71 1 2
4.99 1 2 3	4.99 1 2 3	6.01 1 2 3
4.75 1 2 3 4	4.61 1 2 3 5	5.70 1 2 3 5
4.25 1 2 3 4 5	4.19 1 2 3 5 9	5.33 1 2 3 4 5
3.96 1 2 3 4 5 6	3.55 1 2 3 5 6 9	5.03 1 2 3 5 6 9
4.24 1 2 3 4 5 6 7	3.00 1 2 3 4 5 6 9	4.45 1 2 3 4 5 6 9
4.81 1 2 3 4 5 6 7 8	3.02 1 2 3 4 5 6 9 10	4.83 1 2 3 4 5 6 9 10
3.47 1 2 3 4 5 6 7 8 9	3.11 1 2 3 4 5 6 7 9 10	6.23 1 2 3 4 5 6 7 9 10
3.57 1 2 3 4 5 6 7 8 9 10	3.57 1 2 3 4 5 6 7 9 10 8	8.29 1 2 3 4 5 6 7 8 9 10

[a] The complete set of samples and models made with increasing number of PCs introduced in order of decreasing eigenvalues (usual PCR regression). [b] The complete set of samples for models made with an increasing number of PCs selected according to their best performance. [c] All possible combinations of models for the 15 selected samples only; the best subset of factors is indicated .

scaled scores of few selected samples. The higher cross-validation errors obtained when using only the 15 selected samples (Table 3, column C) can be explained if the 15 selected samples are important for modelling, and when one is deleted, the model decreasing the prediction properties when it is included in the model. It should be noticed that factors 1–6 and 9, which give the model with the lowest RMSECVT, have also been selected as important for the methodology developed here using the is considerably altered. Another explanation could be that the remaining nine samples are not important for modelling, since they are similar to other selected samples. Thus, their deletion does not considerably change the model, which makes the prediction errors smaller.

To show that the proposed methodology, in the great majority of cases, can perform better than the random division of samples into a calibration and an evaluation set, 10,000 models were built by randomly dividing the 24 samples into a calibration set of 15 samples and a test set of nine samples and their RMSEP was evaluated. Only 78 of them gave lower RMSEP values than the model built with the selected samples (Figure 3). This shows that the random division of samples into training and test sets is not a guarantee for building a good quality model, specially if the randomly selected calibration set does not adequately span the experimental domain.

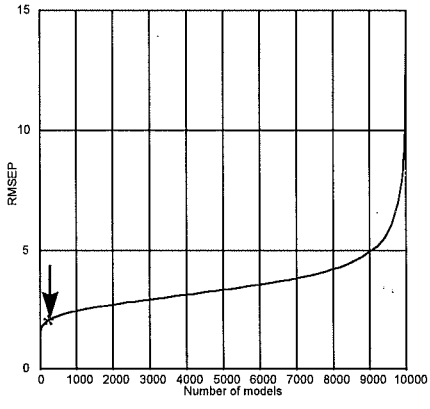


Figure 3. RMSEP of 10 000 models for calcium in water made by randomly dividing the samples into calibration and evaluation sets. The arrow points to the RMSEP of the model built using the methodology proposed.

Data set II. NIR spectra of 46 corn samples. *Selection of the Important Factors To Be Used in the Screening Model.* Initially evaluating the factors according to Malinowski's IND function resulted in selecting 15 factors which were considered to be important. The exchange algorithm searched for the subsets containing N=16-46 samples which maximized the Det($S_{N\times16}{}^T$ $S_{N\times16}$) function. The plot log(Det(M_N)).

UNIVERSITAT ROVIRA I VIRGILI
EXPERIMENTAL DESIGN APPLIED TO THE SELECTION OF SAMPLES AND SENSORS IN MULTIVARIATE CALIBRATION
Joan Ferré Baldrich
ISBN:978-84-691-1875-7/DL: T-337-2008

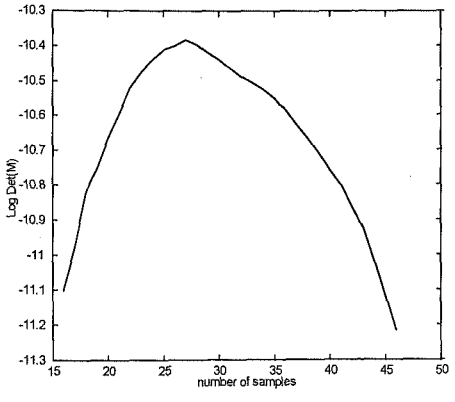


Figure 4. Number of selected samples (N) versus log (Det($\mathbf{M}_N$)). Corn samples, 15 factor model.

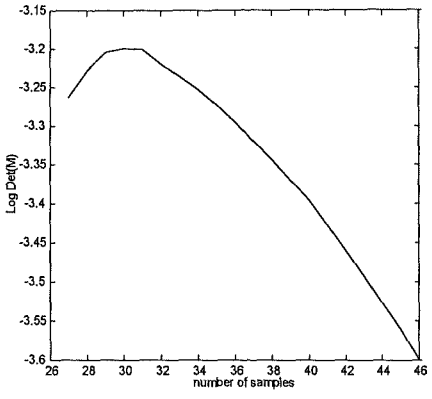


Figure 5. Number of selected samples (M) versus log (Det($\mathbf{M}_M$)). Corn samples, five selected factor model.

Table 4. Moisture of Corn. Regression Results for the Model Made with 15 Factors and the 27 Selected Samples [a]

PC	eigenvalue ($\times 10^{-3}$)	% variance	cumulative % variance	coeff	*t*-test	CV error
1	20.065	92.28	92.28	8.26	48.17	2.91
2	1.617	7.44	99.72	4.85	34.14	0.98
4	0.016	0.07	99.79	1.26	7.31	0.68
3	0.033	0.15	99.94	0.78	5.68	0.46
7	0.000	0.00	99.94	0.54	2.78	0.41
6	0.002	0.01	99.95	0.24	1.17	0.42
9	0.000	0.00	99.95	0.24	1.59	0.42
13	0.000	0.00	99.95	0.23	1.21	0.42
15	0.000	0.00	99.95	0.16	1.10	0.44
10	0.000	0.00	99.95	0.14	0.94	0.42
14	0.000	0.00	99.95	0.10	0.71	0.44
12	0.000	0.00	99.95	0.10	0.57	0.46
11	0.000	0.00	99.95	0.04	0.19	0.48
8	0.000	0.00	99.95	0.03	0.16	0.52
5	0.010	0.05	100.00	0.02	0.11	0.59

[a] The meanings of the columns are the same as for Table 1.

versus number of selected samples N (Figure 4) shows that the subset of 27 samples is the one having the most information per sample. From the experimentally determined moisture content in these selected samples[9], the screening regression model was built. The regression results are listed in Table 4. Only factors 1–4 and 7 are statistically significant according to the t-test (calculated $t_{0.05,11}$=1.80). This agrees with the global minimum error obtained by cross-validation.

Final regression model. Using factors 1- 4 and 7 and the exchange algorithm, the subset of 30 samples had maximum log(Det($\mathbf{M}_M$)) (Figure 5), so only three new samples had to be analyzed for their moisture content to be determined. The regression results for the final model are listed in Table 5. All the factors used are important for prediction as indicated by the t-test ($t_{0.05,24}$= 1.711) the results of which agree with the ones obtained using the minimum RMSECV when the model is built with all the selected factors.

Model validation. The model is accurate according to the F-test, with α=0.398. It should be pointed out that a very good precision value of RMSEP = 0.44 is obtained. Moreover, RMSECV = 0.39 is comparable to RMSECVT = 0.41, indicating that the selected samples cover the experimental domain quite well.

Table 5. Regression Results for the Final PCR Model Made with Factor Numbers 1, 2, 3, 4, and 7 and the 30 Selected Samples [a]

PC	coeff	t-test	CV error
1	8.32	57.27	2.90
2	4.82	41.98	0.92
4	1.31	8.80	0.66
3	0.78	6.85	0.44
7	0.38	2.74	0.39

[a]The meanings of the columns are the same as for Table 1.

Conclusions

A new procedure for selecting the best calibration sample subset for PCR has been developed. It is based on D-optimal design theory and makes use of only the easily obtainable multivariate instrumental responses. The cost in time and effort of the calibration process is substantially reduced because only the selected samples are submitted to chemical analysis by using a reference method. In addition, an approach for quickly selecting the factors with high modeling ability in PCR has been devised. The overall methodology has proven to provide accurate and precise results. The method is of a general nature, and it can be applied to data sets obtained using very different instrumental techniques.

Although only the selected samples are used for calibration, the eigenvector structure is computed for all the instrumental responses available. In this way, all possible causes of variability are taken into account, so ensuring that a representative subset which covers the experimental domain of the chemical constituents is found. However, it should be pointed out that, when few calibration samples are to be used, the quality parameters of the methodology depend very much on the reliability of the analytical results carried out using the reference method.

To sum up, the whole procedure tries to reach a compromise between the quality, in terms of accuracy and precision, that the experimenter demands of the model and the effort, in terms of time and cost, that he or she is ready to put in to build it.

Several research areas related to the present methodology can be developed in the future. The approach developed is being tested by our group to select the minimum number of appropriate samples for multivariate instrument standardization. In addition, since prediction is the main function of the multivariate model, the quality of the predictions furnished by the model, measured by var(c), (G-optimal designs) instead of by minimizing the error of the coefficient estimates, could be a more suitable criterion for selecting the samples to build the final PCR model. This new approach would overcome the distressing problem of not taking into account the errors in the regressor variables to build the multivariate model. However, this would require the development of new sample selection algorithms. Furthermore, a new method to simultaneously designate the calibration and validation samples could be

developed. This methodology would select, from the overall data set, the sample subset that would provide the best estimation of the model coefficients and the samples that give the best indication of the quality of the model.

Acknowledgment

The authors express their gratitude to A. Rius and G. Puchwein for providing the UV-vis and near-IR spectral data, respectively. J. F. thanks the Comissionat per a Universitats i Recerca of Generalitat de Catalunya, for providing a doctoral fellowship (FI/94-7001). Financial support from the Spanish Ministry of Education and Science (DGICyT project BP93-0366) is gratefully acknowledged.

Supporting information available

Mathematical expressions of Fedorov's exchange algorithm. (2 pages). Ordering information is given on any current masthead page.

SUPPLEMENTARY MATERIAL

Fedorov's exchange algorithm[28-29] for selecting D-optimal matrices has three main steps:

1) Initiation: the initial matrix $\mathbf{S}_{N\times(P+1)}^{(0)}$ is built with N randomly selected samples with the condition that the matrix $\mathbf{S}^{(0)T}\mathbf{S}^{(0)}$ must be non-singular. Alternatively, the experimenter can choose the samples it is made up of. In the case of $\mathbf{S}_{M\times(Q+1)}^{(0)}$, only M-N samples are randomly added to the already selected N samples. For simplicity the dimensions are not indicated: $\mathbf{S}^{(0)}$

2) Iteration number *j*:

1. The pair of samples (s^o, s^i) that gives the maximum increase in $Det(S^{(j+1)T}S^{(j+1)})$ is selected. s^o ("out")=$(1, s^o_1, s^o_2,..., s^o_p)^T$ is the sample that leaves the matrix $S^{(j)}$, and s^i("in")=$(1, s_1^i, s_2^i,..., s_p^i)^T$ is one of the candidate samples that enters the matrix.

2. s^o is replaced by s^i in the matrix $S^{(j)}$ so a new matrix $S^{(j+1)}$ is obtained .

3) Stop criterion: the algorithm stops when the increase in $Det(S^{(j+1)T}S^{(j+1)})$ is zero or less than a critical value.

Mathematical expressions:

When a sample s^i enters the initial matrix of experiments $S^{(j)}$ and a sample s^o leaves at the same time, it can be shown [41] that

$$Det(S^{(j+1)T}S^{(j+1)}) = Det\,(S^{(j)T}S^{(j)})(1 + \Delta(s^o,s^i))$$

where: $\Delta(s^o,s^i) = d(s^i) - d(s^o) - d(s^i)\times d(s^o) + [d(s^o,s^i)]^2$ with :

$d(s^o,s^i) = s^{oT}(S^{(j)T}S^{(j)})^{-1}s^i = s^{iT}(S^{(j)T}S^{(j)})^{-1}s^o$

$d(s^i) = d(s^i,s^i)$

$d(s^o) = d(s^o,s^o)$

The maximum increase in $Det(S^{(j)T}S^{(j)})$ is achieved by the pair of samples with largest $\Delta(s^o,s^i)>0$, and this is what the algorithm looks for. The exchanges are faster when only considering the samples with $d(s^i) - d(s^o)>0$ which is a necessary condition for $\Delta(s^o,s^i)>0$. A disadvantage of this algorithm is that the final solution depends on the choice of the initial matrix since it can reach local maxima instead of global maxima. This problem can be solved by repeating the procedure several times with different initial matrices. In our experience, five to ten restarts are enough to find the best solution several times.

Literature cited

1. Beebe, K.R.; Kowalski B.R. *Anal. Chem.* **1987**, *59*, 1007A-1017A.
2. Martens H.; Naes T. *Multivariate Calibration*, Wiley: New York, 1987
3. Kowalski, BR.; Seasholtz M.B. *J. Chemom.* **1991**, *5*, 129-145.
4. Geladi P.; Kowalski B.R. *Anal. Chim. Acta* **1986**, *185*, 1-17.
5. Lorber A; Kowalski B.R. *J. Chemom.* **1988**, *2*, 67-79.
6. Gemperline P.J. *J. Chemom.* **1989**, *3*, 549-568.
7. Aastveit A.H.; Marum P., *Appl. Spectrosc.* **1993**, *47*, 463-469.
8. Naes, T. *J. Chemom.* **1987**, *1*, 121-134.
9. Puchwein G. *Anal. Chem.* **1988**, *60*, 569-573.
10. Schostack, K. J.; Malinowski, E. R. *Chem. Intel. Lab. Syst.* **1989**, 6, 21-29.
11. Skagerberg B.; Bonelli D.; Clementi S.; Cruciani G.; Ebert C. *Quant. Struct.-Act. Relat.* **1989**, *8*, 32-38.
12. Norinder U.; Högberg T. *Acta Chemica Scand.* **1992**, 24, 363-366.
13. Hellberg S.; Sjöström M.; Skagerberg B.; Wold S. *J. Med. Chem.* **1987**, *30*, 1126-113536.
14. Wold S.; Sjöström M.; Carlson R.; Torbjörn L.; Hellberg S.; SkagerbergB.; Wikström C.; Öhman J. *Anal. Chim. Acta* **1986** *191* 17-32.
15. Carlson R. *Design and optimization in organic synthesis.* Elsevier: The Netherlands 1992
16. Zemroch, Peter J. *Technometrics* **1986**, *28*, 39-49.
17. Honigs D.E.; Hietfje G.M.; Mark H.L. ; Hirschfeld T.B. *Anal. Chem.* **1985** 57, 2299-2303.
18. Marengo E. ; Todeschini R. *Chem. Intel. Lab. Syst.* **1992**, *16*, 37-44.
19. Kalivas, J. *J. Chemom.* **1991**, *5*, 37-48.
20. Hruscha W.; Norris, K. *Appl. Spectrosc.* **1982**, *36*, 261-265.
21. Isaksoon T.; Naes T. *Appl. Spectrosc.* **1990**, *44*, 1152-1158.
22. Naes T.; Isaksoon T. *NIR news* **1994** ,*5*, 16-17
23. Box, G.E.P. ; Hunter, W.G.; Hunter, J.S. *Statistics for Experimenters* Wiley: N.Y. 1978
24. Nishii, Ryuei. *Discrete Mathematics* **1993**, *116*, 209-225.
25. Steinberg D.V.; Hunter W.G. *Technometrics* **1984**, *26*, 71-130
26. Kennard, R.W. ; Stone, L.A. *Technometrics* **1969** ,*11*, 137- 148
27. Atkinson, A.C. ; Donev.A.N. *Optimum Experimental Designs.* Oxford Statistical

Science Publications: Oxford 1992

28. Fedorov, V.V. *Theory of optimal experiments.*(translated and edited by W.J. Studden and E.M. Klimko) Academic Press: New York, 1972
29. Mathieu D. *Contribution de la Méthodologie de la Recherche Experimentale à l'étude des relations Structure-Activité.* Thèse Sciences. Marseille, 1981
30. Sutter, J.M.; Kalivas J.H. ; Lang P.M. *J. Chemom.* **1992**, *6*, 217-225.
31. Sun, J. *J. Chemom* **1995**, *9*, 21-29.
32. Malinowski, E. R. *Factor Analysis in Chemistry*, 2nd de; John Wiley & Sons Inc: New York 1991
33. Wold S., *Technometrics* **1978**, *20*, 397-405.
34. Eastment H.T.; Krzanowski W.J. *Technometrics* **1982**, 24, 73-77
35. Mandel, J. *J. Res. of the National Bureau of Standards* **1985**, *90*, 465-476.
36. Khuri, A.I.; Cornell, J.A *Response Surfaces. Designs and Analyses*; Marcel Dekker: New York, 1987; pp 36-38.
37. Mathieu, D.; Phan-Tan-Luu, R. NEMROD ver. 3.0. L.P.R.A.I. - Université d'Aix-Marseille 1995
38. Mandel, J.; Linning, F. J. *Anal. Chem.* **1957**, *29*, 743-49
39. Rius A.; Callao M.P.; Rius F.X. *Anal. Chim. Acta*, in press.
40. Matlab. The Mathworks, South Natick, MA, 1994.

3.3 Determination of ethylene content in poly(propylene-ethylene) copolymers using near-infrared spectra (NIR) and multivariate calibration

(submitted)

Villagrasa C., Ferré J., Larrechi M.S., Rius F.X., García C* [1].

Department of Chemistry, Universitat Rovira i Virgili. Pl. Tarraco, 1. 43005 Tarragona.
[1] *Transformadora de Polipropileno TDP. PO Box 1175. Tarragona. Spain*

A new method for determining the ethylene content in poly(propylene-ethylene) copolymers using near-infrared spectra (NIR) in the 1666-1767 nm range and multivariate calibration is discussed. Three multivariate calibration methods were studied; principal component regression (PCR), principal component regression selecting the factors according to its predictive ability (PCRSF), and partial least-squares regression (PLS). PLS was found to have the best precision. The absence of bias in this model was assessed by performing the joint statistical test of the slope and the ordinate in the regression of $c_{pred\text{-}cv}$ versus c_{known} for the calibration samples taking, into account the errors in both axes.

1.Introduction

Heterophasic poly(propylene-ethylene) copolymers are widely used for industrial purposes because of their extreme toughness. This property is usually measured by the impact strength,[1,2] which is related, among other variables, to the ethylene content. Determining the percentage of ethylene is an important analysis in the quality control process of these plastics.

In industry, ethylene concentration is usually determined by using the infrared (IR) spectra of the pressed films of these copolymers[3,4]. The area of the bands between 750 and 690 cm^{-1}, which corresponds to the absorption of the methylenic sequences, is used to calculate the ethylene concentration. This area, however, is previously divided by the area of the bands between 4361 and 3950 cm^{-1} to correct for the thickness of the film. Univariate linear calibration of these data is then carried out using internal standards as a reference. The variability coefficient of this analysis is between 5-10%. It is sometimes difficult to accurately determine ethylene concentration because of the presence of talc which produces a band in the IR spectra that overlaps the ethylene band (Figure 1), thus making univariate calibration unsuitable in this case.

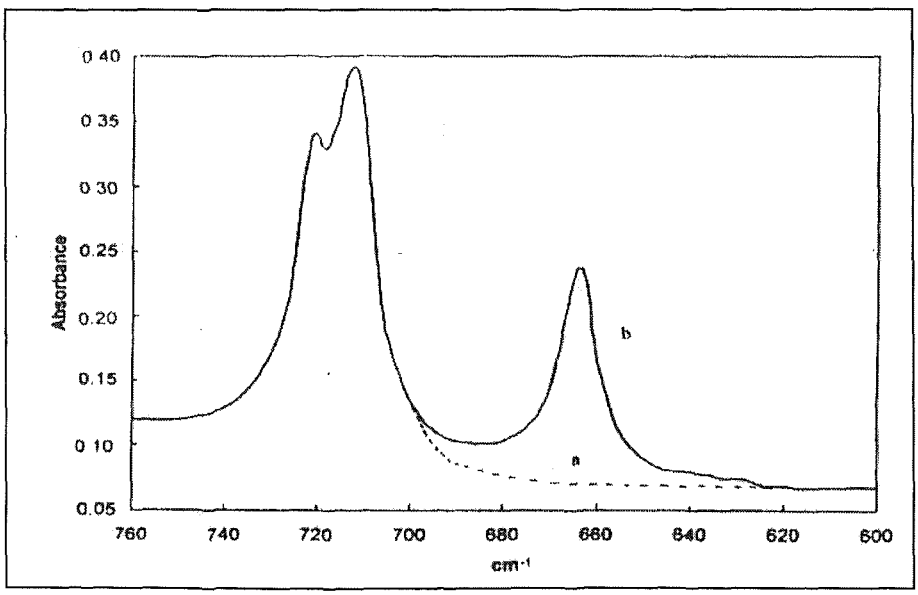


Figure 1. A typical IR absorption spectrum of poly(propylene-ethylene) copolymer a)without talc b)with talc

In this study, an alternative method for determining ethylene in poly(propylene-ethylene) copolymers is described. It uses the near-infrared (NIR) spectra of the samples between 1666 and 1767 nm which correspond to the first overtone of CH stretching bands[5]. Although the use of NIR spectroscopy to determine structural properties in different types of plastics such as polyethylene and polyurethane has been reported[6,7] to our knowledge, neither quantitative nor qualitative application of poly(propylene-ethylene) copolymers have been described. Multivariate calibration must be applied to quantify the ethylene concentration since there are no selective wavelengths. Three multivariate calibration methods were studied and compared: principal component regression (PCR), PCR selecting factors for their prediction ability (PCRSF) and partial least-squares regression (PLS).

2. Theoretical background

Notation. Matrices are represented by bold capital letters, column vectors by bold lower-case letters and scalars by italic characters. The superscript ^{T} means transposed. The subindices in a matrix indicate its dimensions. Let $\mathbf{R}_{I\times J}$ be the column mean-centered matrix of instrumental response data for I samples and J sensors and $\mathbf{c}_{I\times 1}$ the vector of the ethylene concentration in the I calibration samples.

Principal component regression. In PCR, $\mathbf{R}_{I\times J}$ is decomposed according to:

$$\mathbf{R}_{I\times J} = \mathbf{T}_{I\times P}\mathbf{P}_{J\times P}^{T} + \mathbf{E}_{I\times J} \qquad (1)$$

where the columns in $\mathbf{T}_{I\times P}$ are $P \leq min(I,J)$ uncorrelated underlying factors, $\mathbf{P}_{J\times P}$ is the loading matrix and $\mathbf{E}_{I\times J}$ is a matrix of residuals. After determining which Q principal components are important for regression, $\mathbf{c}_{I\times 1}$ is regressed versus $\mathbf{T}_{I\times Q}$ according to:

$$\mathbf{c}_{I\times 1} = \mathbf{T}_{I\times(Q+1)}\,\beta_{(Q+1)\times 1} + \mathbf{f}_{I\times 1} \qquad (2)$$

where $\beta_{(Q+1)\times 1}$ is the vector of regression coefficients, a column vector of 1's has been appended to $\mathbf{T}_{I\times Q}$ to account for a constant term and the elements of $\mathbf{f}_{I\times 1}$ are the calibration error terms. An estimate of β can be found using the least-squares method[8]. The Q factors that provide the best predictive model can be found by cross-validation[9] of several regression models built with a different number of factors. These factors are put into the model depending on the value of their corresponding eigenvalue.

Principal component regression with selection of factors (PCRSF). It has been shown that the factors with the largest eigenvalues do not necessarily give the best predictive PCR model[10]. An alternative PCR model can be built by considering only the most-predictive principal components, which can be selected with a screening method. The procedure used, which has the advantage of being very fast, is as follows:

(1) $\mathbf{R}_{I\times J}$ is decomposed according to equation (1). The number of factors in $\mathbf{T}_{I\times P}$ is selected on the basis of the optimal number of factors used to build the 'usual' PCR model and it is overdetermined to ensure that all factors containing information are taken into account.

(2) The Range-Midrange Transformation[11] is performed on $\mathbf{T}_{I\times P}=\{t_{ip}\}$ for every sample $i=1,\ldots,I$ and every factor $p=1,\ldots,P$ so as to obtain $\mathbf{S}_{I\times P}=\{s_{ip}\}$:

$$s_{ip}=(t_{ip}-C_p)/R_p \qquad (3)$$

where:

$$C_p = [\max(t_{ip}) + \min(t_{ip})] / 2$$
$$R_p = [\max(t_{ip}) - \min(t_{ip})] / 2$$

with $\max(t_{ip})$ and $\min(t_{ip})$ being the highest and lowest score values for the pth factor. This scaling technique makes the scaled scores s_{ip} for every factor span the range [–1,+1].